

Introduction to Statistics for Community College Students

(First Edition)

By Matt Teachout

**College of the Canyons
Santa Clarita, CA, USA**



*This chapter is from Introduction to Statistics for Community College Students,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Special thanks to all of the people that made this book possible.

I would like to thank all of the math and statistics teachers at College of the Canyons. Your work to improve statistics education continues to inspire educational reform.

I would also like to thank James Glapa-Grossklag, Brian Weston and the College of the Canyons OER staff for their tireless work to support free high quality materials for students through the OER movement.

Special thanks to Randy Ades, Gevork Demirchyan, Kathy Kubo, Joe Gerda, Udani Ranasinghe and Rupa Sinha for supporting me through this project.

I would also like to thank the fall 2019 pilot team: Alvard Adamyan, Randy Ades, Steve Brown, Roy Erickson, Jason Moss, Said Najafi, Linda Newland, Eric Pham, Udani Ranasinghe, Amar Singh, Rupa Sinha and Juliana Yankey. Your feedback and suggestions were fabulous.

Most especially, I would like to thank my wife Link for her love, support, and patience with me during the endless hours of writing.



This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Introduction

We live in a world of big data. Our children grow up with technology, computers, the internet, and a vast amount of information at their fingertips. How can we make sense of this vast amount of information? How can we know when we are being deceived or if data is biased? How can businesses or hospitals make sound decisions based on data? The answer to all of these questions lies in the study of statistics and data science.

The study of statistics and data science is vital in our modern age. Statisticians, data scientists and data analysts are now in high demand. Every company, hospital, sports team or college needs trained employees who can collect and analyze data and help make good decisions based on data. In the U.S., there is a huge deficit in the number of people trained in this area. Not only do we not have enough statisticians, data scientists and data analysts, but also we do not have enough statistics teachers and statistics tutors.

Statistics is a deep well of knowledge that men and women have devoted their lives to studying. In this book, I will attempt to give you some useful tools and an overarching picture of statistics, but we will only be playing in a puddle compared to that deep well of statistical knowledge. I do hope to whet your appetite though and encourage you to study statistics and data science past this initial class. Maybe we can add a few more statisticians, data scientists and data analysts to our ranks. The world desperately needs you.

Vocabulary

In many ways, studying statistics is like learning a new language. Statisticians and data scientists are experts in collecting and analyzing data. Yet they also need to be able to explain their findings to a world that has very little understanding of statistical reasoning. Statistical terms often have a very different meaning than we would find in a dictionary. New statistics students often find the vocabulary overwhelming. For this reason, you will find definitions of important terms at the beginning of each section. It is important to read these carefully and commit them to memory. Many of my past students found it very helpful to write each term and definition on a 3x5 card. They would then study their 3x5 cards throughout the semester. Here are some terms and definitions to get you started.

Data: Information in all forms.

Statistics: The science of collecting, preparing and analyzing data.

Statistician: An expert in the science of statistics. The leaders in the field.

Data Mining: The process of collecting and storing data.

Data Science: Many feel this is just another word for statistics, but with maybe less emphasis on high-level statistical analysis and more emphasis on using computer science to collect and format data.

Data Scientist: A specialist in data science.

Data Analyst: A specialist in analyzing data to make good decisions.



*This chapter is from **Introduction to Statistics for Community College Students**,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Notes about OER and Creative Commons Licensing

There are many fabulous books on statistics and analyzing data. Unfortunately, they are extremely expensive and many college students cannot afford the cost. I wrote this book to help people learn to analyze data. It is free to use the material in this book, update it, add to it, print it or just read it. It is an open educational resource (OER) and so anyone can use it.

Many college students struggle to balance work and family with their education. One of the biggest roadblocks for many students is the cost of textbooks. Many students simply cannot afford the cost of textbooks. They attend classes without purchasing books and materials needed for the class. Of course, this is a major impediment to passing their classes, but the students have no choice. They simply cannot afford \$100-\$200 textbooks. For this reason, I believe strongly in open educational resources (OER). Open source materials like this book are available and are virtually free for students.

This textbook is licensed through Creative Commons as “Attribution CC-BY”. Creative Commons describes this license as follows: “This license lets others distribute, remix, tweak, and build upon (the author’s) work, even commercially, as long as they (give) credit (to the author) for the original creation.” This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.” If you need to see the license deed or legal code, please go to <https://creativecommons.org/licenses/> and look under the “CC-By” section.

Supplementary Materials

Links to the individual sections and problem sets can be found at www.matt-teachout.org. Just click on the “statistics” tab and pick what chapter you wish to study. (Posted summer 2019)

There are also additional supplementary materials under the “statistics” tab at www.matt-teachout.org. (Posted summer 2019)

- Teaching notes for instructors.
- Data sets.
- Affective domain assignments.

The following supplementary materials may be found under the “Instructor Resources” and “Statistics Resources” tabs at www.matt-teachout.org. (Posted summer 2019)

- Suggested teaching schedules.
- Example projects.
- Practice Problem Answer Keys.



*This chapter is from **Introduction to Statistics for Community College Students**,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Important Note about Technology

We live in the age of computers, internet and a huge volume of data. No practicing statistician or data scientist uses a calculator or tables to analyze data. You cannot even begin to analyze a data set of 100,000 values by hand with a calculator. You need high-powered computer software. There are many statistics software programs on the market, but very few of them are free.

If you read the history of statistics, you will find brilliant scientists, mathematicians and people in business who were trying to figure out data, but had no access to a computer. (Computers had not been invented yet.) Our pioneers of statistics dreamed of the day that they could compute statistics and graphs and analyze data with the touch of a button. They invented complicated techniques for analyzing data because they had no choice. Today, computers can calculate statistics and graphs directly.

Here is the problem. Many statistics classes taught in high schools, community colleges and even some universities are teaching statistics as if computers have not been invented yet. They are teaching the techniques developed by our pioneers of statistics before the computer age. This is a terrible approach to the subject, especially for the thousands of students that actually want to work in the field. A statistics class should be a study of how to practically collect and analyze data with a computer, not a class on what to do if computers had not been invented yet.

Are formulas important in statistics? Yes. We look at formulas to understand what they tell us about the data and the world around us. The pioneers of statistics did an amazing job of addressing the major ideas of statistics with formulas and inventive calculations. However, we should not use a formula and a calculator to calculate a statistic for a data set with 10,000 values or use charts that list critical values and P-values. High-powered computers with statistics software can calculate the statistic and make graphs directly. Then students can focus on the analysis part, and explore and discover the meaning behind the data.

This book will show students how to use statistics software to calculate statistics and graphs. I want students to learn to explore and analyze the data and not spend all their time just trying to calculate something. Remember, no one pays a data scientist to calculate something a computer can already do. A data scientist is paid to explore the data and explain what the data may be telling us. The key question is not "how is this calculated?" The key question is "what can I learn from this data?" Computers are tools to explore the data.

StatKey

Teaching statistics with computer software is very important, but many schools and students cannot afford to pay for software. For this reason, I prefer to use software that is free for students, but is also relatively easy to use. My favorite software by far is StatKey. StatKey is hosted online at www.lock5stat.com. StatKey is free, well organized, easy to use, and has a nice blend of modern and traditional statistics capabilities.

Statcato

Another free software that I sometimes use is Statcato. Statcato is a traditional JAVA based software found at www.statcato.org. It can be saved to your computer or it can be opened with the "Java Web Start" function. Some students have difficulty getting Statcato to open on their home computers. For this reason, the homework sets will already have Statcato printouts. That way, if students have any technical issues, they can still complete the homework.

Other Programs

You can of course use the book with any statistics software. Most basic statistics software programs are very similar to Statcato, but may not be free for students. Instructors that use a program other than StatKey or Statcato will need to give software directions to their students.



*This chapter is from **Introduction to Statistics for Community College Students**,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Data Sets

The national (GAISE) guidelines for teaching statistics recommend that you use real data. Allowing students to learn statistics principles through analysis of real data is key. With that being said, there are many places where raw data can be found and used. The key data sets throughout this book are located at www.matt-teachout.org under “Statistics” and “Data Sets”. It is wise to save the book and the data sets on your computer. That way you have them when you need them and do not have to keep going to the website.

The Computer Dilemma

A statistics or pre-statistics class should be taught in a computer lab. It is important to allow the computers to do the difficult calculations. Students need to focus on interpretation and discovering the meaning behind the data. They cannot do that if they spend all their time trying to calculate with a formula or making graphs by hand.

If your school wants to teach statistics or pre-statistics, but you cannot teach in a computer lab, here are some suggestions for you.

1. Reserve unused computer labs. Some schools may have a couple computer labs that are not always in use. Schedule your statistics and pre-statistics classes in order to use the computer lab. Even if you can only reserve the lab once a week or once every two weeks, it will be a huge help to students.
2. Have groups of students share computers. If you do have a few computers in your classroom, you can divide the class up into groups and share computers. This has many advantages like encouraging explanations to one another and teamwork.
3. Teachers can use their own computer or laptop to project statistics software on a screen and have the class analyze the data with you. Teachers without any computer can make printed copies of the software printouts for your class and for exams. It is a poor substitute for a computer lab, but it is much better than teaching statistics as if computers have not been invented.

Organization of Chapters and Sections

Many statistics books organize the material into four or five large units with several chapters in each unit. For example, a book may have five units with thirty chapters and give one exam for each unit. I prefer to organize the material into four large chapters with several sections in each chapter. This book has twenty-eight sections broken into four chapters. A good rule of thumb is to plan on one major exam for each chapter.

Pedagogy

Instructors often ask me about pedagogy and the difficulty in teaching statistics to underprepared students. I thought I would share my core principles that define how I like to teach statistics classes.

1. Active Learning and Classwork

- It is vital for students to practice problem solving and using the statistics software during class. When teaching a new section, I like to give a short lecture that focuses on vocabulary, understanding the main ideas, and using technology. Then I have the students work on the problems in the section with my help. They can work in groups or individually, but they need to work with the data themselves and get used to explaining concepts to others. I then give them a break and repeat the process (short lecture and work on problems).



*This chapter is from **Introduction to Statistics for Community College Students**, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

2. Alternative Approaches to Homework

- I try to minimize the number of problems students do outside of class and maximize the number of problems done in class with me to tutor and help them. Homework usually consists of them finishing the problems they did not finish during class. I also like assigning alternative homework assignments that focusses on main concept understanding and explaining. Small writing assignments work well this. For example, I like to have students write a paragraph explaining the main concepts they learned during class. It is also a way for me to check for understanding. Statistical concepts can be very difficult, so I also like to supplement my students' knowledge for homework. For example, I have the students watch a video online, take notes and turn in their notes to me.

3. Use Technology

- In a way, I see my statistics classes as job training. No working statistician uses a calculator and a formula especially with big data. The job of a working data scientist is to collect data, use computer software to explore the data, make good data driven decisions, and be able to explain it to others. In my class, I like to mimic that process. I do not spend time teaching students how to calculate something by hand with a formula. Instead, I focus on using and interpreting the computer software. The computer software does the calculations. The students have the more important job of exploring, understanding and explaining.
- I tend not to front-load when teaching technology. I like to teach the theory and show the software together. It is difficult to find a program that does everything, especially one that is free. In this book, I used a blend of Excel, Statcato and StatKey. Data sets are managed in Excel spreadsheets.

4. Understanding over Calculating

- Formulas are very important in statistics, but not to calculate. As I have said, no one working in the field calculates statistics by hand with a formula and calculator. We use computer software to calculate. I like to study the formulas with my students. These formulas are programmed into statistical software like Statcato and StatKey. We need to understand what the formula is calculating and what it tell us about the data. What are the limitations of this formula? When can we use it and when can we not use it? It is about understanding the theory behind the formula.

5. Blend Traditional and Modern Statistics

- Computer technology drastically changed the landscape of statistics. While it is important for student to understand traditional formulas and techniques, I have found that modern computer-based techniques like bootstrapping and randomization have benefits also, especially in conceptual understanding. I like to blend the traditional and modern statistics techniques. In this book, I use Statcato for traditional calculations. I use StatKey for data exploration and exposing students to more modern computer-based techniques.

6. Vocabulary

- Statistics is like learning a new language. Statistical terms have their own definitions that are rarely the same as the one in the dictionary. Students need to learn to explain statistical ideas and understand statistical reasoning. I constantly cover and review the key terms and ideas in each section. Without understanding the vocabulary, students will really struggle with the theory.



Table of Contents

Introduction to Statistics for Community College Students (1st edition)

by Matt Teachout

Introduction & Table of Contents

Chapter 1: Collecting and Analyzing Data

Section 1A – Two Types of Data

Section 1B – Collecting Data

Section 1C – Bias

Section 1D – Experimental Design

Section 1E – Categorical Data Analysis

Section 1F – Normal Quantitative Data Analysis

Section 1G – Non-normal Quantitative Data Analysis & Summary Statistics

Chapter 1 Review

Chapter 2: Estimating Population Parameters

Section 2A – Statistics and Parameters

Section 2B – Sampling Variability & Sampling Distributions

Section 2C – The Central Limit Theorem

Section 2D – Introduction to Confidence Intervals

Section 2E – One-Population Mean & Proportion Confidence Intervals

Section 2F – Two-Population Mean & Proportion Confidence Intervals

Section 2G – One-Population Variance & Standard Deviation Confidence Intervals (*Optional*)

Chapter 2 Review

Chapter 3: Introduction to Hypothesis Testing

Section 3A – Null & Alternative Hypothesis

Section 3B – Test statistics & Critical Values

Section 3C – P-value & Significance Levels

Section 3D – Conclusions

Section 3E – Type 1 & Type 2 Errors

Section 3F – One-population Mean & Proportion Hypothesis Tests

Chapter 3 Review

Chapter 4 – Relationship Tests

Section 4A – Categorical/Quantitative Relationships: Two-Population Mean Hypothesis Test

Section 4B – Categorical/Quantitative Relationships: Analysis of Variance (ANOVA)

Section 4C – Proportion Relationship Tests: Two-Population Proportion Hypothesis Test

Section 4D – Proportion Relationship Tests: Goodness of Fit Test

Section 4E – Categorical Relationships: Contingency Tables

Section 4F – Categorical Relationships: Categorical Association Test

Section 4G – Quantitative Relationships: Correlation & Regression

Section 4H – Quantitative Relationships: Correlation Test (*Optional*)

Chapter 4 Review



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Chapter 1: Collecting and Analyzing Data

Vocabulary

Data: Information in all forms.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not represent the population.

Introduction: The goal of collecting and analyzing data is to understand the world around us. How data is collected is very important. The goal of collecting data is to get “unbiased” data that represents the population. Analyzing biased data may result in incorrect conclusions and lead to a misguided view of the world around us. It is also important to have a goal in mind when you collect data. Are we trying to find a population percentage from categorical data or a population average from quantitative data? Are we trying to show that two variables are related or are we trying to show cause and effect? Data needs to be collected differently depending on what goal you have in mind.

Section 1A – Two Types of Data – Categorical and Quantitative

One of the most important factors when analyzing data is to determine what type of data you have and how many variables you are analyzing. Let us start with the type of data.

There are two general types of data, categorical and quantitative.

Categorical Data

Categorical data (or qualitative data) are generally labels that tell us something about the people or objects in the data set. For example, what country do they live in, what is the person's occupation, or what kind of pet they have?

Usually categorical data is made up of words (do you smoke - yes or no), but occasionally a number can be used as a category. For example, a zip code can be used instead of the place a person lives. The numbers “1” and “2” may be used instead of yes and no.

Quantitative Data

Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. For example: a list of people's heights in inches, or their weights in kilograms, or a list of how many dogs are there in various animal shelters across Los Angeles. Notice in each of these cases the data is numerical and an average seems appropriate in the context. We can find the average height, the average weight, or the average number of dogs in animal shelters in Los Angeles.

Numbers used as categories

Remember, not all numeric data is quantitative. Ask yourself if the numbers are measuring or counting something and if an average would make sense. For example, a list of people's zip codes are numbers but an average zip code would not really tell us anything. In addition, identity numbers like hospital ID numbers, student ID numbers or social security numbers are not measuring anything and an average would not make sense in the context so they are not quantitative.



Practice Problems Section 1A

1. The following spreadsheet can be found on the website www.matt-teachout.org. Just click on the “statistics” tab and then “data sets”. This data was taken from bears. Use the bear data to classify each column of data as categorical or quantitative. If the data is quantitative, what are the units? If the data is categorical, indicate how many different options there are in that category.

AGE (months)	Month Data Taken	Gender	Head Length (in)	Head Width (in)	Neck Circum (in)	Length (in)	Chest (in)	Weight (Lbs)
19	July	male	11	5.5	16	53	26	80
55	July	male	16.5	9	28	67.5	45	344
81	September	male	15.5	8	31	72	54	416
115	July	male	17	10	31.5	72	49	348
104	August	female	15.5	6.5	22	62	35	166
100	April	female	13	7	21	70	41	220
56	July	male	15	7.5	26.5	73.5	41	262
51	April	male	13.5	8	27	68.5	49	360
57	September	female	13.5	7	20	64	38	204
53	May	female	12.5	6	18	58	31	144
68	August	male	16	9	29	73	44	332
8	August	male	9	4.5	13	37	19	34
44	August	female	12.5	4.5	10.5	63	32	140
32	August	male	14	5	21.5	67	37	180
20	August	female	11.5	5	17.5	52	29	105
32	August	male	13	8	21.5	59	33	166
45	September	male	13.5	7	24	64	39	204
9	September	female	9	4.5	12	36	19	26
21	September	male	13	6	19	59	30	120
177	September	male	16	9.5	30	72	48	436
57	September	female	12.5	5	19	57.5	32	125
81	September	female	13	5	20	61	33	132
21	September	male	13	5	17	54	28	90
9	September	male	10	4	13	40	23	40
45	September	male	16	6	24	63	42	220
9	September	male	10	4	13.5	43	23	46
33	September	male	13.5	6	22	66.5	34	154
57	September	female	13	5.5	17.5	60.5	31	116
45	September	female	13	6.5	21	60	34.5	182
21	September	male	14.5	5.5	20	61	34	150
10	October	male	9.5	4.5	16	40	26	65
82	October	female	13.5	6.5	28	64	48	356
70	October	female	14.5	6.5	26	65	48	316
10	October	male	11	5	17	49	29	94
10	October	male	11.5	5	17	47	29.5	86
34	October	male	13	7	21	59	35	150
34	October	male	16.5	6.5	27	72	44.5	270
34	October	male	14	5.5	24	65	39	202
58	October	female	13.5	6.5	21.5	63	40	202
58	October	male	15.5	7	28	70.5	50	365
11	November	male	11.5	6	16.5	48	31	79
23	November	male	12	6.5	19	50	38	148
70	October	male	15.5	7	28	76.5	55	446
11	November	female	9	5	15	46	27	62
83	November	female	14.5	7	23	61.5	44	236
35	November	male	13.5	8.5	23	63.5	44	212
16	April	male	10	4	15.5	48	26	60
16	April	male	10	5	15	41	26	64
17	May	male	11.5	5	17	53	30.5	114
17	May	female	11.5	5	15	52.5	28	76
17	May	female	11	4.5	13	46	23	48
8	August	female	10	4.5	10	43.5	24	29
83	November	male	15.5	8	30.5	75	54	514
18	June	male	12.5	8.5	18	57.3	32.8	140



2. The following spreadsheet can be found on the website www.matt-teachout.org. Just click on the “statistics” tab and then “data sets”. This data was taken from various cereals. Use the cereal data to classify each column of data as categorical or quantitative. If the data is quantitative, what are the units? If the data is categorical, indicate how many different options there are in that category.

Name	Manufacturer	Target (Adult or Child)	Shelf displayed in store	Calories per serving	Carbs (grams per serving)	Fat (grams per serving)	Fiber (grams per serving)	Potassium (milligrams per serving)	Protein (grams per serving)	Sodium (milligrams per serving)	Sugar (grams per serving)	Vitamin (Percent of Daily need per serving)	Consumer Report Magazine Rating	Serving Size (Cups per serving)	Weight (Ounces per serving)
Captain Crunch	Quaker	Child	Middle	120	12	2	0	35	1	220	12	25	39	0.75	1
Cocoa Puffs	General	Child	Middle	110	12	1	0	35	1	180	13	25	23	1	1
Flit	General	Child	Middle	110	13	1	0	35	1	140	12	25	28	1	1
Apple Jacks	Kellogg	Child	Middle	110	14	0	1	30	2	125	14	25	33	1	1
Corn Chex	Kellogg	Adult	Bottom	110	12	0	0	25	2	180	3	25	41	1	1
Corn Flakes	Kellogg	Adult	Bottom	100	11	0	1	35	2	260	2	25	46	1	1
Nut & Honey	Kellogg	Adult	Middle	120	15	1	0	40	2	190	9	25	30	0.67	1
Stacks	Kellogg	Child	Middle	110	9	1	1	40	2	70	15	25	31	0.75	1
Nutri-Grain	General	Adult	Bottom	100	15	1	2	90	2	220	6	25	40	1	1
Cracklin	Kellogg	Adult	Top	110	10	3	4	160	3	140	7	25	40	0.5	1
Grape-Nuts	Post	Adult	Top	110	17	0	3	90	3	170	3	25	53	0.25	1
Honey Nut	General	Child	Bottom	110	11.5	1	1.5	90	3	250	10	25	31	0.75	1
Nutri-Grain	Kellogg	Adult	Top	140	21	2	3	150	3	220	7	25	41	0.67	1.33
Product 19	Kellogg	Adult	Top	100	10	0	1	45	3	320	3	100	42	1	1
Total Raisin	General	Adult	Top	140	15	1	4	230	3	190	14	100	29	1	1.5
Wheat Chex	Kellogg	Adult	Bottom	100	17	1	3	115	3	230	3	25	50	0.67	1
Cornmeal	General	Adult	Top	130	13.5	2	1.5	120	3	170	10	25	30	0.5	1.25
Life	Quaker	Child	Middle	100	12	2	2	95	4	150	6	25	45	0.67	1
Whego	America	Adult	Middle	100	16	1	0	95	4	0	3	25	55	1	1
Quaker Oats	Quaker	Adult	Top	100	14	1	2	110	4	135	6	25	50	0.5	1
Wheat R	Kellogg	Adult	Top	150	16	3	3	170	4	150	11	25	34	1	1
Quaker Oatmeal	Quaker	Adult	Bottom	100	14	2	2.7	110	5	120	0	0	51	0.67	1
Chewies	General	Child	Bottom	110	17	2	2	105	6	290	1	25	51	1.25	1
Special K	Kellogg	Adult	Bottom	110	16	0	1	55	6	290	3	25	53	1	1

3. Determine if each of the following variables are quantitative or categorical.

- The number of milligrams of Aspirin given to heart attack patients.
- The various types of cars being sold at a used car lot.
- Determining if a person smokes marijuana or not.
- The number of bicycles sold at various bicycle stores in Seattle, WA.
- The types of birds observed in Florida.
- The number of grams of gold found in various streams across northern California.
- The various types of cardio classes offered at gyms across Los Angeles, CA.
- The number of cardio classes offered at gyms across Los Angeles, CA.
- The city a person lives in.
- The amount of money in peoples' bank accounts.
- The various zip codes from addresses at a post office.
- The drivers' license numbers from various taxi drivers.
- The number of taxis driven in New York City on various days of the week.



Section 1B – Collecting Data

Vocabulary

Population: The collection of all people or objects you want to study.

Census: Collecting data from everyone in the population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not reflect the population.

Random: When everyone in the population has a chance to be included in the sample.

One of the most important goals in data science is to learn about the world around us (populations). It is very difficult to understand populations sometimes because data may be biased and not reflect the population very well. Bias can occur in many different ways, but certain ways people collect data have more bias than others do. Using a method for collecting data that increases bias is sometimes called “sampling bias”.

It is important to be aware of various methods used to collect data, the good and the bad.

Method 1: Census

A census is the best way to collect data if it is possible. If our goal is to learn about the population, it makes sense to collect data from everyone in the population. There are ways for a census to be biased, but in terms of the collecting method, a census is the best. Unfortunately, it is almost impossible to collect a census if your population is large. Most statisticians and data scientists are only able to collect a sample, data collected from a small subgroup of the population.

Method 2: Simple Random Sample

If a statistician or data scientist cannot collect a census, the preferred method is to collect a random sample. A random sample is one where everyone in the population has a chance to be in the sample, so it tends to represent the population better than other non-random samples. It is nowhere near as good as a census, but as I said, a census is usually not possible.

A simple random sample is one where individuals in the population are selected randomly. This can be a difficult process. The usual method is to assign everyone in a population a number and then use a random number generator in a computer program to pick random numbers. Computer programs have many built in randomization functions for this purpose. If you have a spreadsheet of the entire population, a computer can also randomly select individuals from the list. The key with a “simple random sample” is that you are selecting people or objects one at a time. Collecting data randomly and one at a time gives greater flexibility to your sample. Almost any grouping is possible with a simple random sample, so it tends to represent populations better than other samples.

There are many examples of a simple random sample. Many statistics companies use a random phone number generator that randomly gives phone numbers. They then call the phone numbers randomly chosen and try to get information from people that answer the phone. The U.S. government may have a computer randomly select social security numbers to select individuals for a sample. A company may have a computer randomly select employee ID numbers to select individuals for a sample.



Method 3: Convenience Sample

People often find collecting a census or a simple random sample difficult, so they chose to collect data in whatever way seems easiest. A sample collected this way is often called a “convenience sample” and is popular with people not trained in statistics. A convenience sample usually has much more bias than a random sample and may not represent the population very well.

An example of a convenience sample is collecting data from your friends and family. This is fine if your population of interest is your friends and family, but will by no means represent a large population. Another example might be standing outside of a store or post office and collecting data from people that leave the store. Beginning statistics students may walk into a mall and collect data from whomever they bump into. They mistakenly think that these are random samples, but they are not. A random sample means everyone in the population has a chance to be included in the sample. Not everyone in the population has a chance to bump into you at a mall or come out of a store at 2:30 pm on a Tuesday afternoon. These are convenience samples and generally do not reflect the population very well.

Method 4: Voluntary Response Sample

Some say that all surveys are bad, but that is not the case. A survey is just a form to collect data from people. When a company takes a census of all its employees, it may require all of the employees to fill out a survey. That is a census. As long as no other forms of bias creep into the data, a census will probably be a very good representation of the population. The point is that giving a survey is not the issue. The issue is whom you give the survey to and who is allowed to fill out the survey.

A voluntary response sample puts a survey out into the world and allow anyone to respond. The usual method used today is to put a survey on a website and allow anyone that comes across the survey to answer. The survey can also be a mailed to every address in a given population. Again, those that fill it out self-select themselves to be in our data.

On the surface, a voluntary response sample may seem like a good way of collecting data. It usually gives a large amount of data. Does this really allow everyone in the population a chance to answer? It turns out the answer is no. Ask yourself the following question. When you are surfing the web and a survey pops up, do you fill it out? I have been asking my statistics classes that question for years and rarely have anyone that says that they do fill out surveys. The key problem is that only certain types of people will fill out a survey voluntarily. It may be a person who is bored and has nothing better to do. It is certainly not a person with three children, working a full time job and going to college full time. It may also be a person who is upset by or feels very passionate about the topic in the voluntary response survey. They are so upset by the lack of pay for teachers that they are willing to fill out a survey to tell you what they think. The point is that voluntary response surveys tend to over-sample people that are bored or upset and under-sample everyone else. For this reason, voluntary response samples can be very biased and may not represent the population very well.

I have had many students ask me if sample size is important. Isn't a voluntary response sample of five thousand people better than a random sample of fifty people?" I would tell them that though sample size is important, method is important also. The voluntary response sample of five thousand would tend to over-represent people that are bored or upset about the topic. It does not represent typical people in the population. The random sample of fifty people, while a small sample size, at least does not have that bias.



Method 5: Cluster Sample

A cluster sample is one where data is collected from groups of people in a population instead of one at a time. For example, a company that has 250 stores worldwide might have a computer randomly select ten stores and get data from those people that work at those ten stores. Notice this would be a random sample since every employee has a chance to be in the data. If their store was chosen, then they will be included in the sample. This is not a simple random sample however, since they are not choosing one at a time. This example is sometimes called a “random cluster sample”. While it is a good method for collecting data, it has less flexibility than a simple random sample. Think of it this way. In a simple random sample, any grouping is possible, but in this random cluster example, only groups of people that work at the same store can be chosen. It is still a random sample though, and would tend to be more representative of the population than non-random samples like convenience or voluntary response.

It is good to note that the goal of a cluster sample should be to choose the groups of people randomly. If we choose groups of people that are convenient to collect data from, our cluster sample will have more sampling bias and will not represent the population nearly as well.

Method 6: Stratified Sample

One of the most common studies done in statistics is to compare groups. We may compare data from 2016 to data from this year. We may compare people living in Canada to people living in Australia. To compare groups, you need to collect a stratified sample.

Some people in statistics explain a stratified sample as comparing two or more groups in one population. I like to think of it as comparing two or more populations. Whether you explain a stratified sample as comparing groups in one population or comparing populations, the key is that you are comparing.

For example, we may want to compare the percentage of adults in the U.S. with diabetes to the percentage of children in the U.S. with diabetes. Some statistics authors think of this as comparing adults and children in the one population of all people in the U.S. I like to think of it as comparing the population of U.S. adults to the population of U.S. children.

Another example may be to compare the mean average salary of people working in London, England to the mean average salary of people working in Toronto, Canada. Again, a stratified sample is needed because we are comparing.

To do a stratified sample, we often take a simple random sample from each group. I like to think of it as taking a simple random sample from each population you want to compare. In the previous example, we may collect a simple random sample of adults in the U.S. and another simple random sample of children in the U.S. We then can calculate the sample percentages that are diabetic from each sample and use statistical methods to compare them. For the salary example, we can collect a simple random sample of salaries for people working in London, and another simple random sample for people working in Toronto. The goal is then to use statistical methods to compare the mean average salaries.

It should be noted that when taking a stratified sample, we should use randomization. Again, if we just take a convenience sample from each group or voluntary response sample from each group, we will likely have a lot more bias and the data will not reflect the population (or populations) as well as we would like.

Many people confuse a cluster sample with a stratified sample because they both involve groups. The goal of a cluster is to get data on and analyze one population, not to compare. You are just collecting data from groups of people from that one population instead of one at a time. The goal of a stratified sample is to compare two or more populations so we need to collect data from each population.



Method 7: Systematic Sample

A systematic sample is one where we use a system to collect the sample. Usually it involves collecting data from every fifth person that comes in your store or every twentieth person on a list.

For example, let us suppose we want to collect a sample of students from our college. We could look at an alphabetical list of the names of all students that attend our college and then chose every 50th person on the list. Is this a random data set? Ask yourself this question. Does everyone on this list have a chance to be chosen? No. Only the 50th, 100th, 150th, 200th and so forth have a chance. People from 1-49 have no chance. People from 51-99 have no chance. Therefore, it is not a random sample. This may not be random, but we may make the argument that it is representative of the population. This method would have less bias than convenience or voluntary response samples. There is a way to incorporate randomization into the method. Many data scientists have a computer chose a random number between 1 and 50. Suppose it is 33. Then they collect data from the 33rd person on the alphabetical list. Now, from there use the system of choosing every 50th person. Therefore, they would choose the 33rd person, then the 83rd person, then the 133rd person and so on. Making the first choice random, makes the whole data set random, because everyone on the list now has a chance to be chosen.

Summary

So let us summarize the various methods.

- An unbiased census is the best way to collect data to represent a population, because we are collecting data from everyone in the population.
- If you cannot do a census, then use a random sample of some sort. It may be a simple random sample, random cluster, or a random systematic sample. The main thing is that if you are collecting a sample, randomization needs to be involved.
- Voluntary response samples and convenience samples tend to be very biased and should be avoided if possible.

Practice Problems Section 1B

Directions: For each of the following, identify the population of interest. Then identify the method used to collect the data (census, systematic, convenience, voluntary response, cluster, stratified, or simple random). Explain why you chose your answer and if the method will represent the population of interest or not?

1. The admissions department at a college wants to see how many of their students would be in favor of using a new program to register for classes. They put a link on their website so that any students that want to try out the program can. The students can then take a survey and say how well they like the new system.
2. Rick works for a sports equipment manufacturing company. He wants to compare the opinion of his older employees to the new employees. To do this, he separates all the employees into two groups, employees that have been with company five or more years and those that have been with the company less than five years. He then chooses 12 of his most trusted older employees and 16 new employees that have proven themselves and ask what they think about changing the medical insurance coverage.
3. Michelle, a teacher at a local high school, wants to see how many students at her high school will be attending community college. She gives the students in her one section of advanced placement U.S. History a questionnaire to fill out that asks where they will be attending college.
4. Jamie is working at the Republican recruiting committee in her city. She is curious how many people that live in her city will vote for the Republican candidate in the next election. She uses a computer to randomly select phone numbers in her city. She then calls those phone numbers to ask people about their voting preferences.



5. Rachael works at the Democrat recruiting center in her hometown. To determine what percent of people will vote for the Democratic candidate, she obtains a list of all residents in her town and decides to ask every 40th person on the list.
 6. Laya is passionate about bringing an NFL football team to her city. She needs to take an opinion poll about how people in her city would feel about raising taxes in order to build a stadium for a professional football team. She randomly selects 75 streets in her city and asks every person living on those streets.
 7. Micah is the CEO of large software development company. He wants to see if his employees have any ideas about areas of software development that the company should pursue. He has every single employee in his company fill out a questionnaire outlining his or her ideas. He gives the employees a stipend on their paycheck to pay them for their time it took to fill out the questionnaire.
 8. Tara wants to collect data on people living in Portland Oregon. She wants to know how many cups a coffee they drink per day. She went to a few supermarkets close to her house and asked people as they were leaving the store.
 9. Julius works for a company in Toronto, Canada that manufactures eyeglasses. He wants to know what styles of glasses people in Toronto prefer. He randomly selects phone numbers in Toronto and calls them to ask about glasses preference.
 10. Hugo works at a public library and wants to collect data on all of the people that come to the library. He looks up every single person in the library database and notes the number of books that he or she has checked out in the last six months.
 11. A company is designing a new type of smart phone. They want to know how much memory people prefer in their smart phones. They put a question up on several search engines and allow anyone to answer.
 12. A college wants to collect data on their students to see how often they use the various student services offered by the college. They randomly select 60 classes and collect data from all of the students taking those classes.
 13. A clothing store is designing a new line of athletic wear. They want to compare the percentage of teenagers that prefer the new line of athletic wear to the percentage of adults that prefer the new line of athletic wear. They take a random sample of teenagers and ask them about the new athletic wear. Then they take a random sample of adults and ask them about the new athletic wear.
 14. Brian is collecting data for his statistics class project on the amount of time people spend on social media per day. He asks people in his college classes and at his church how many minutes they spend on social media per day.
 15. A store that sells BBQ's in North Carolina wants to know what percentage of people own a "smoker BBQ". They ask every third person that enters the store if they own a smoker BBQ or not.
-



Section 1C – Bias

Vocabulary

Population: The collection of all people or objects you want to study.

Bias: When data does not reflect the population.

The purpose of collecting data is to learn about the world around us, to learn about populations. The problem is that many people that collect data may not have had any training in Statistics or Data Science. The result is that many data sets collected do not reflect the population very well. When this happens, we say that the data is biased.

Many people think that if you collect a random sample or a census, it will guarantee that you will have an unbiased data set. This is not true. There are many types of bias and it is possible to have a census or a random sample that does not reflect the population very well. It is critical that we be aware of these other forms of bias and to try our best to make sure they are not incorporated into our data sets.

Sampling Bias

In the last section, we said that the best way to collect data is a census. This means that we collected data from everyone in the population. If we cannot collect a census then we should try to collect a random sample or at least a sample that represents the population. We said that convenience samples or voluntary response samples are inherently biased and usually do not reflect populations very well. Using a bad data collecting method like convenience or voluntary response gives rise to sampling bias. When sampling bias occurs, it usually means the technique for collecting the data was poor.

Question Bias

It has been said that there are lies, bad lies, and then there is statistics. There is some truth in this. People with specific agendas may twist data and statistical analysis to suit their purpose. One way to do this is question bias.

A question bias occurs when someone phrases a question in a specific way to force people to answer the way they want.

For example, suppose a politician wants to show that most people in her city agree with her policy on raising taxes to improve health care. She may collect a great simple random sample, but ask the question this way.

“Health care in our city is extremely bad. Hospitals and urgent cares are in bad need of renovation and need better supplies. The elderly need to know that we have not forgotten them. We need to improve the quality of care for our children. Will you support my policy for improving health care across our city?”

Phrasing the question this way, no one would guess that the real issue was whether to raise taxes. People, hearing this question, think about helping the children and elderly, not about taxes. When a large percentage of people answer that they support her plan, she now has data to support her agenda.

When you collect data, you want to ask questions in a neutral way that does not attempt to sway people in one direction or another. It also should not leave out key information like what the real question is. If the politician had simply asked people in the simple random sample if they would be in favor of raising taxes to improve health care, she likely would have gotten a much smaller percentage of people to agree.

Notice that in this example, the data was a simple random sample. This is a good data collection method, as methods go. However, the incorporation of a question bias into the data makes the data very bad. This simple random sample does not reflect the population at all. The data has been manipulated to support an agenda.



Response Bias

Many topics are very difficult to get data on because people do not feel comfortable answering truthfully. If you ask people if they are addicted to alcohol or drugs, they are likely to deny it even if they do struggle with substance addiction. People may lie about their age, weight, or salary. When a large percentage of people in your data lie, you have a response bias in your data.

Suppose a church wants to collect data on how many hours per week their congregation spends helping the homeless. They decide to have every person in their congregation fill out a survey listing how many hours per week they help the homeless. Remember a census is usually the best way to collect data about a population, but this census has a problem. It is a topic that people are likely to lie about. People may put a higher number of hours on the survey than they really do so that they will not look bad to the church leaders. The average number of hours calculated from this data will likely be larger than the population average number of hours. Even though this is a census, it probably does not reflect the population very well.

When dealing with topics that people are likely to lie about, the data scientist needs to have a plan to deal with the response bias. Instead of asking people their weights, maybe they weigh them on a scale. Instead of asking people about their salary, maybe they look at paycheck stubs. Instead of asking people about substance abuse, they may collect data from agencies that support people with addiction.

Deliberate Bias

We have stated already that people may misuse statistics and data in order to support their agenda. Deliberate bias is another example of this. Deliberate bias can take on a variety of forms. It could be someone deliberately leaving out groups from the data. The most common is collecting data and then leaving out the data of people that disagreed with you. It can also be deliberately lying about the results of the data report. Maybe the data makes your restaurant or hospital or school look bad, so people just falsify their records and deliberately lie about the results of the study. The data may be census or a random sample but the conclusions have been falsified and the data distorted.

Deliberate bias is a major problem in statistics. It is also a good reason to have an independent statistics company collect the data and do the analysis. Use a statistics company that is not tied to the government, business, hospital, restaurant or politician in question. An independent statistics company is less likely to lie about the results or to falsify the data, though it is naive to think that it never happens.

I tend to be suspicious about internal statistics reports that come out where the company, government or politician refuses to share the data. We are supposed to take their word for it and agree with the findings. There are good reasons why companies do not share data, but I always wonder if they are they afraid that someone analyzing that data would come to a very different conclusion?

There is large worldwide discussion of ethics for people that work in the fields of statistics or data science. Statistical analysis is a powerful tool and is a vital discipline to understand and improve the world around us, but falsifying records or manipulating data should never be an option. It is not only unethical, but also makes people question the integrity of our science.

Sometimes specific groups in the population may not be represented very well in the data. This also falls under the umbrella of deliberate bias. For example, suppose a person may wish to collect data on adults living in a city. However, they only collected data from people living in the wealthier areas of that city. It may not have been done deliberately. It could just be that the person collecting the data did not think about certain groups in the population that are not being represented. In large cities, the homeless are often difficult to get data on. A person collecting data has to have a plan for getting data that will represent all the groups in their population, including the homeless.



Non-response Bias

Non-response bias is becoming a huge problem for all people that collect data. A computer may randomly select people to collect data from, but more often than not, the person does not want to participate. They may fear identity theft or are just too busy to participate. It is a huge problem. We need data. We need to understand the world around us, but it now becoming increasingly difficult to get unbiased data. Many people that collect data report that sometimes only one in every five randomly selected people will participate and give data. The problem of non-response bias continues to get worse. This makes us consider what type of person gives data and if that person is truly reflective of all people in the population.

To combat the problem of non-response bias, many people that collect data offer a reward system for people that will participate and give data. This may help a little, but then offering a reward may incorporate its own bias into the data.

Summary

There are many reasons why data may not reflect a population. It is a mistake to think that a random sample or a census will always be devoid of bias. It is increasingly important to be aware of possible sources of bias and to strive to keep them out of our data as much as possible. The goal of data collecting is to collect unbiased data that reflects the population. Always phrase questions in a neutral way that avoids question bias. Have a plan for collecting data about topics where people are likely to lie. We have to have a good plan on how we will collect data. It should be a census or a random sample, but we should also think about groups that may not be represented. We need to avoid deliberate bias and never falsify reports or distort data to support someone's agenda.

Practice Problems Section 1C

1. Define each of the following and give an example of each.

- | | |
|------------------|----------------------|
| a) Population | f) Response Bias |
| b) Census | g) Sampling Bias |
| c) Sample | h) Deliberate Bias |
| d) Bias | i) Non-response Bias |
| e) Question Bias | |

Directions for #2-10: For each of the following scenarios, describe the population of interest and all of the types of bias that the data may have (Question, Response, Sampling, Deliberate or Non-response). There may be more than one type of bias involved. Explain your answers and if there is bias, what groups of people were not represented.

2. We are interested in calculating the percent of children in LA County that are up to date with their vaccines. To figure this out, a person put a survey up on the yahoo webpage asking the following question: "Is your child up to date with their vaccines?" The computer will keep track of the number of people that answer yes or no.
3. We are interested in finding what percent of people in the U.S. agree or disagree with vaccinating children. To figure this out, we randomly selected 350 people in the U.S. and asked them the following question: "In order to save children from devastating diseases, do you agree that all children should be vaccinated?"
4. We are interested in finding out how many people in the U.S. have had whooping cough this year. To figure this out, we called every major hospital in the United States and asked how many people at their hospital were diagnosed with whooping cough this year.
5. We are interested in finding out what percent of Americans use Cocaine. We randomly chose 400 Americans and asked them if they use Cocaine or not.



6. What is the average age of college students in Canada? Since my cousin lives in Canada, I asked him to drive to two colleges near his house and ask people he bumps into what their age is.
 7. Julie is interested in calculating the yearly income of adults in Palmdale. She drives around Palmdale, stops at certain streets, and then asks people that live on that street what their yearly income is? She skips streets that look "sketchy" as she is worried about her safety.
 8. A college wants to collect data on their students to see how often they use the health office for mental health counseling. They randomly select 35 classes and collect data from all of the students taking those classes. They asked the following question. "It is very important for all college students to have mental health support. College students report having depression, anxiety and high stress levels. The college offers free mental health counseling at the health office. Have you taken advantage of these mental health services?"
 9. A pharmaceutical company took random samples of their pills to check that the pill has the correct type and amount of medicine. They noticed that several of their pills did not have the correct amount of medicine, but decided to delete this data.
 10. An auto manufacturer wants to collect data on the type and number of mechanical problems in their cars. They decide to keep data only on all cars brought to their dealerships nationwide.
-

Section 1D – Experimental Design

Vocabulary

Explanatory Variable: The independent or treatment variable. In an experiment, this is the variable causes the effect.

Response Variable: The dependent variable. In an experiment this the variable that measures the effect.

Confounding Variables (or lurking variables): Other variables that might influence the response variable other than the explanatory variable being studied.

Experimental Design: A scientific method for controlling confounding variables and proving cause and effect.

Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

Placebo: A fake medicine or fake treatment used to control the placebo effect.

In statistics, we often want to determine if there is a relationship or association between two variables. We also may want to measure the strength of the relationship. For example, we may want to know if there is a relationship between blood pressure and heart rate. We may want to see if living in tropical climates is associated with having nut allergies.

In order to show that two variables are related or associated we use an observational study. We would collect data and use statistical methods to analyze and measure the strength of the relationship. However, showing that two variables are related does not prove that one causes the other.

Association ≠ Causation!!!!

Why?

Let us suppose that we have shown that there is a strong relationship between drinking alcohol and getting into a car accident. This tells us that alcohol consumption is an important factor to be considered when studying car accidents.



*This chapter is from **Introduction to Statistics for Community College Students**, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

However, this does not prove that drinking alcohol causes car accidents. Many factors go into having a car accident besides how much alcohol they consume. Can you name a few?

Other factors that may influence having a car accident besides alcohol: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like texting, eating or changing a radio station), using drugs, ...

These are called “confounding variables”. Confounding variables are factors that might influence your response variable other than the explanatory variable you are studying. In this case, factors that might influence having a car accident other than how much alcohol the driver consumed. Some statistics books call these “confounding variables” or “lurking variables”.

Note: The explanatory variable (alcohol consumption) is not a confounding variable. Alcohol is the explanatory variable we were studying. Confounding variables are factors other than alcohol that might influence the response (car accident).

Here is the point. If many variables were involved in having a car accident, it would be wrong to say that the alcohol was solely responsible for the car accident. Alcohol is just one of many factors involved. We have shown that drinking alcohol is related but we have not proven cause and effect. To prove cause and effect we need to deal with the confounding variables.

Experimental Design

So how do we prove cause and effect? It is difficult. You would need to prove that each confounding variables is not involved and so it is only the explanatory variable that is causing the response. The key is controlling the confounding variables. Thankfully, scientists have put a great deal of thought into this process of controlling confounding variables and proving cause and effect. We call this process “experimental design”.

Experimental design is a scientific method for controlling confounding variables and proving cause and effect. A key component to experimental design is the creation of similar groups through random assignment.

To control confounding variables, we will need to create two or more groups of people or objects that are very alike. One way to do this is by “random assignment”. Random assignment is a process where you take a group of people or objects and randomly split them into two or more groups. The randomly assigned groups tend to be very similar. If we do not think the groups are similar enough, we can use techniques like blocking or direct control to make the groups even more alike.

Another way to make alike groups is to use the same group of people twice. Think about it. The two groups would be perfectly alike. They would have the same ages, same amount of stress, same genetics, same blood pressures and the same jobs.

Example

Let us look at the previous example. How do we prove that drinking alcohol does cause car accidents?

Explanatory (treatment) Variable: Drinking alcohol or not

Response Variable (what we will measure): Did the person get into a car accident or not?

So how do we set up an experiment to prove that drinking alcohol causes car accidents? The first thing is to list out your possible confounding variables.

Possible Confounding Variables: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like passengers, texting, eating or changing a radio station), other drugs, gender, race, genetics

To control the confounding variables, we need to create two groups of people. The two groups should be the same (or at least as similar as possible) in all areas that the confounding variables address. Therefore, the groups should



have similar ages, similar driving experience, similar cars and car condition, similar road conditions and similar distractions, similar genders, similar race and ethnicity, similar genetics and reflexes.

There are two ways to go about this. Let us suppose we have a group of 80 adult paid volunteers to conduct this experiment. One option would be to randomly put the volunteers into two groups and try to make the groups as similar as possible. A better option in this case would be to use the same people twice.

We had the people in the experiment drive an obstacle course sober. They must have no alcohol or other drugs in their system. They all used the same car on the same track with the same weather. The course was designed with cones and we will monitor how many cones the people hit. They all were not allowed to have any other person in the car. There was no other distractions as radios and phones were not allowed. We will monitor how many car accidents they had by checking how many cones they hit.

Now we will have all the people drink a certain amount of alcohol and then drive the course again. It is important to see that the alcohol (treatment) group was made up of exactly the same people as the sober (control) group. The response variable we measured was the number of cones they hit.

Conclusion

The results found that the alcohol group hit significantly more cones (significantly more car accidents) than the sober group. We have now proven that drinking alcohol causes car accidents.

Think about it. It cannot be the ages of the drivers or driving experience. The two groups had the exact same ages and the exact same driving experience. It cannot be gender, race, genetics, or reflexes. The two groups had the exact same genders, race, genetics, and reflexes. It cannot be drugs or other distractions like phones or radios. Neither group had drugs or any other distractions. If you notice, every one of the confounding variables is the same in the two groups. The only difference was that one group had alcohol and the other did not. Therefore, the only reason why the alcohol group had significantly more accidents is the alcohol. The experiment has proven that drinking alcohol causes car accidents.

Note: It is easy to confuse the two variables in an experiment with the two groups. They are not the same thing.

In this case, the explanatory variable is having alcohol or not. The response variable is the number of cones (accidents) the drivers had. The two groups are decided by those that have explanatory variable (alcohol) and those that do not. In this case, the two groups are the exact same people measured twice.

We usually call the group that has the explanatory variable the “treatment group” and the group that does not have the explanatory variable the “control group”.

Example 2

When a pharmaceutical company needs to prove that a medicine works, they must use experimental design. In the United States, pharmaceutical companies have to prove to the Food and Drug Administration (FDA) that their medicine has the effect it is supposed to and is relatively safe with few side effects.

Suppose a company has a new blood pressure medicine on the market and needs to prove to the FDA that taking it does decrease a person's blood pressure. The company needs to prove cause and effect.

If we have to prove cause and effect, we need an experiment. The first step is to think about the possible confounding variables. What are some reasons why a person's blood pressure might decrease other than taking this new medicine?

Possible Confounding Variables? Stress, Diet, Exercise, Genetics, Age, Gender, Race, Genetics, taking other medicines ...

To set up the experiment we need to create two groups of people that are similar in these areas. We start with a group of volunteers with high blood pressure that want to try out this new medicine. We randomly assign the people into two groups. Amazingly when scientists randomly assign people into two groups, the groups tend to be a lot alike. The two groups would have similar numbers of people in each race, similar number of males and females, similar



numbers of stressed out people, similar numbers of people that exercise a lot or do not exercise. The people running the experiment can also exercise direct control and intentionally assign people to certain groups to make the groups even more alike.

Human Brain (placebo effect)

There is a problem with our experiment. If a person believes something is true, their brain can tell the body to manifest physical responses. We call this the “placebo effect”. Think of it this way. The group that thinks they are getting blood pressure medicine will not be as stressed out about it and their blood pressure may decrease slightly because of that belief. Similarly, the group that thinks they are not getting blood pressure medicine will be more stressed and worried and their blood pressure may increase because of that belief. In a sense, the human brain is a confounding variable that we need to control.

Placebo (fake medicine)

To control the placebo effect as a confounding variable, we need the groups to believe the same thing. One group cannot think they are getting medicine, and the other group cannot believe they are not getting medicine. So we introduce a placebo or fake medicine. The treatment group gets the real blood pressure medicine and the control group gets a fake medicine (placebo). No one in the experiment knows if he or she will be receiving real medicine or a placebo. Some may ask, “Won’t that make them more stressed and increase their blood pressure?” Yes. The key is that the two groups will be equally stressed and believe the same thing. That way we control the placebo effect.

For this to work, the people in the experiment cannot know if they are getting the medicine or a placebo. This is called “single blind”. When scientists first started using placebos, they were shocked to find that the people in the experiments somehow knew if it was a placebo. This defeated the whole purpose. It turned out they could tell by the body language of the person giving the medicine. The person giving the medicine tended to act differently if they were giving the real medicine versus a placebo. So the standard for an experiment about medicines is to use a “double blind” approach. A double blind experiment means that neither the people in the experiment, nor the people giving the medicine, know if it is a placebo or not. Someone knows though. The scientists keep very careful track of who receives a placebo and who receives the medicine. The person directly giving the medicine or placebo cannot know if it is a placebo or not.

Double blind works well. The people in the experiment no longer know if they are receiving a placebo or the real medicine. The experimental design has controlled the placebo effect.

Conclusion

Since we have controlled all of the confounding variables, the experiment has the possibility of proving cause and effect. We still need to see the blood pressures of both groups and make a conclusion. If the treatment group had a significantly lower average blood pressure than the control group, this would prove that taking the medicine does cause a person to have lower blood pressure. If the treatment group and control group have relatively the same average blood pressure, then we may conclude that the medicine is not effective in lowering blood pressure. This would be bad news for the pharmaceutical company. Deciding if one group is significantly higher than another can be very difficult. We will study confidence intervals, test statistics and P-value in later chapters to address this.

Summary

Use an experiment to control confounding variables and prove cause and effect. The groups in the experiment should be the same people either measured multiple times or separated by random assignment. The main idea is that the groups should be very similar in all areas that involve confounding variables. Experiments with medicines should be double blind with a placebo to control the placebo effect.

Use an observational study to see if there is a relationship (association) between two things. Remember observational studies do not control confounding variables, so cannot prove cause and effect.

How can I tell if a study is an experiment or not? Generally, look for random assignment. An experiment usually does not have a random sample of people from the population. The people in the experiment are usually volunteer. The volunteers are then randomly assigned into two or more groups. Random assignment means that they are not trying to apply something to the population, but instead are trying to use experimental design in order to prove cause



and effect. If a study takes a random sample from the population, but does not randomly assign, it is probably just an observational study and cannot prove cause and effect.

Note: It should be noted that there are more complex forms of experiments than the types listed in this section. It may not be possible to randomly assign people into two groups. In that case, the scientist need to prove that each confounding variable is not involved. That is a more complex case that you may see in more advanced statistics classes.

Practice Problems Section 1D

Ruler Experiment Directions: Divide class into groups of three or four. Each group will need a ruler and their cell phones. It is best to stand up during this activity. **Procedure:** Student A will hold the cell phone in their dominant hand and then hold their non-dominant hand straight out in front of them with their hand curved. The fingers should not be very close or very far away from the thumb. While student A is texting, student B holds the bottom of the ruler up inside of student A's non-dominant hand. Student B should hold the ruler from below student A's hand. The top of student A's hand should be about 5 inches on the ruler. Student B releases the ruler and student A tries to catch it. Student C records the number of inches on the top of the ruler before caught. Student C will take the catch length, subtract off the 5 inches, and then record the difference. If student A misses the ruler all together, then student C will just put "drop". Each student should attempt to catch the ruler while texting three times. Then repeat the process, but this time the students will attempt to catch the ruler with their non-dominant hand without a cell phone. Continue until all students have done the experiment three times without the cell phone and three times with the cell phones. Alternate the person releasing the ruler and the time before released. Collect the data for the "with phone" catches and drops in one column. In another column, collect the data for the "no phone" catches. When done, give the data to the instructor. Put the without cell phone/with cell phone data up on the board without names. The instructor or a student will collate the following results for the whole class: the mean average catch length with the cell, the mean average catch length without the cell, the total number of drops with the cell, the total number of drops without the cell.

Use your class data to answer the following questions as group. If you were absent on the day your class did the ruler experiment use the following data.

Ruler Experiment Data (Previous Class)

	With Phone	No Phone
Mean Average Catch (inches)	10.3 inches	8.2 inches
Number of Drops	41 drops	7 drops

1. What is the explanatory (treatment) variable? What was the response variable?
2. Why did we bother to have the person catch the yardstick without the phone?
Wouldn't it of been quicker to just record the catching with the cell phone?
3. What were the two groups of people in the experiment? Were they alike?
Why didn't we randomly assign the groups?
4. What are some of the confounding variables in this experiment?
What are some steps that we took to control these variables?
5. Was this experiment blind, double blind, or neither? How do you know?
6. What did the class data show? Does texting cause slow reflexes?
How do you think this experiment might apply to driving while texting?



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

(#7-11) Define the following terms and give an example of each.

7. Observational Study
8. Experiment
9. Explanatory Variable
10. Response Variable
11. Confounding Variables
12. Random Assignment
13. Placebo
14. Placebo Effect
15. Single Blind
16. Double Blind

(#17-21) Directions: Determine if each of the following studies are an observational study or an experiment. Explain why. Can the study prove cause and effect or just a relationship? Why? If the study is an experiment, list some confounding variables that need to be controlled.

17. Dramamine is a common medication used in preventing and treating nausea, vomiting and dizziness caused by motion sickness. This medication has become a staple for thousands of people who travel by boat, car or plane. We need to prove that Dramamine is effective in preventing and treating the symptoms of motion sickness. Volunteers were randomly assigned into two groups. One group received Dramamine and the other received a placebo. The amount of motion was the same for all of the people. They were then asked to rank their motion sickness on a scale of 1 to 10.

18. Unemployment has become a very important topic in the United States and worldwide. We wish to understand how unemployment may be related to the tax rate. To shed light on this issue, we took a random sample of countries around the world and compared the average tax rate to the unemployment rate.

19. Tuberculosis (TB) is a disease that affects millions of people worldwide. TB is a contagious bacterial infection that affects the lungs. Doctors have long speculated that the percentage of people with Tuberculosis is higher in low income, crowded cities. A medical study was done to see if there is a relationship between low income, crowded cities and a high percentage of people with Tuberculosis. They took a random sample of cities and collected data about the size and the number of people. They then compared it to the number of cases of tuberculosis.

20. College students in the United States have long claimed that listening to music while studying causes them to retain information at a higher rate. We want to prove that this is not true. Listening to music while studying does not cause a person to retain information at a higher rate. We took a group of volunteer college students and randomly put them into three groups. The people in each group had to memorize the same information. They were ranked as high retention or low retention. One group had to listen to their favorite music, another group had to listen to a music they hated, and the third group had no music at all. The volume of music was the same for all of the people.

21. A study was done to determine if there is an association between obesity and diabetes. Obesity and diabetes data was taken from a random sample of adults.



Introduction to Categorical & Quantitative Data Analysis

Vocabulary

Data: Information in all forms.

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A population value, which is sometimes calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be. For example, a population mean average or a population percentage.

Introduction

We learned that, in order to learn about the world around us, we need to collect and analyze data. Our goal is to understand populations. Sometimes we can collect data from everyone in the population (census) and sometimes we can only collect data from a small subgroup of the population (sample). Either way, once we have the data, we need to be able to analyze it. This chapter focuses on the basics of data analysis. If you remember, there are two types of data, quantitative (numerical measurements) and categorical (labels). We analyze quantitative data very differently than categorical data, so it is always vital to ask yourself a couple key questions.

- Was the data collected correctly, either an unbiased census or an unbiased large random sample?
- Is the data quantitative or categorical?
- Is their one data set or are we trying to analyze relationships between two data sets?

We will learn about rules for judging sample sizes in the next few chapters. This chapter focuses on being able to analyze the sample data or census data you have.

When analyzing data we rely on numbers calculated from the data that can help us understand the key features of the data set. If these numbers were calculated from a sample, they are called statistics. If these numbers are calculated from an unbiased census, they are called parameters. Most of the time, we only have sample data, so it is vital to understand and explain statistics.

Note on calculation: We live in the age of “big data”. No one today calculates statistics by hand, especially for a data set of ten-thousand values. Even a sample of one-hundred can be overwhelming to calculate. Statisticians and data scientists rely on computers to calculate statistics. The focus should be on understanding the meaning and correct use of the statistic, not on calculating by hand with a calculator.



*This chapter is from **Introduction to Statistics for Community College Students**,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Section 1E – Categorical Data Analysis

Vocabulary

Percentage (%): An amount out of 100. For example if 72 out of every one-hundred employees opts to use a company's HMO insurance, we would say that 72% of the employees are using the HMO insurance.

Proportion: The decimal equivalent of a percentage. To calculate, divide the percentage by 100 and remove the percent symbol.

Proportion and Percentage Conversions

To analyze categorical data, we focus on exploring various types of percentages and compare them. In statistics, the decimal equivalent to a percentage is often called a "proportion".

To convert a decimal proportion into a percentage, we multiply the proportion by 100%. This moves the decimal point two places to the right. Do not forget to add the % symbol.

Example: Convert 0.047 into a percentage.

$$0.047 \times 100\% = 4.7\%$$

To convert a percentage into a decimal proportion, we divide by 100 and remove the percentage symbol. This moves the decimal two places to the left. Do not forget to remove the % symbol.

Example: Convert 52.9% into a decimal proportion.

$$52.9\% = 52.9 \div 100 = 0.529$$

Calculating Proportions and Percentages from Categorical Data

In order to calculate a decimal proportion from categorical data, you will need to find the amount (count, frequency) and divide by the total.

$$\text{Decimal Proportion} = \frac{\text{Amount (Frequency)}}{\text{Total}}$$

Counting how many people share a certain characteristic or even a total number of cars in a data set can take a long time in a big data set, however technology can help. Statistics software can count much quicker and easily than we can. In this section, we will assume we know the amount and the total.

Suppose a health clinic has seen 326 people in the last month and 41 of them had the flu. If we were analyzing their data, the first thing we would like to do is find what proportion of the patients have the flu. It is not a difficult calculation and can be done with a small calculator.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687$$

Should we round the answer? Proportions and Percentages are usually rounded to the three significant figures. Proportions are usually rounded to the thousandths place (3rd place to the right of the decimal).

Let us review rounding. We want to round the above answer to the thousandths place, which is the "5". Always look at the number to the right of the place value you are rounding. If the number to the right is 5-9, round up (add 1 to the place value). If the number is 0-4, round down (leave the place value alone). After rounding cut off the rest of the decimals.



Therefore, in the previous answer we want to round to the thousandths place (5). The number to the right of the 5 is a 7. So should we round up or down? If you said round up, you are correct. Therefore, we will add 1 to the place value and the 5 becomes a 6. Now we cut off the rest of the decimal and our approximate answer is 0.126.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687 \approx 0.126$$

Decimal proportions are vital in the analysis of categorical data, but many people have trouble understanding the implications of a decimal proportion like 0.126. That is why we often convert the proportion into a percentage.

How to convert a decimal proportion into a percentage

To convert a decimal proportion into a percentage, multiply by 100 and put on the “%” symbol. Think of it like taking 100% of the decimal proportion. When you multiply by 100, the decimal moves two places to the right. Some people prefer to move the decimal, but I find students make fewer errors when they just multiply by 100 with their calculator.

$$\text{Percentage} = \text{Decimal Proportion} \times 100\%$$

Look at our previous example of the number of cases of the flu at a health clinic. We used the amount and total to calculate the decimal proportion.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687 \approx 0.126$$

So what percentage of the patients had the flu? All we need to do is multiply the decimal proportion 0.126 by 100% to get the percentage equivalent.

$$\text{Percentage} = \text{Decimal Proportion} \times 100\% = 0.126 \times 100\% = 12.6\%$$

So 12.6% of the patients at the health clinic were seen for the flu. This can be alarming information to the health clinic if that is an unusually high percentage.

Notice that the percentage still has three significant figures, but is rounded to the tenths place (one place to the right of the decimal). Rounding to the tenth of a percent is a common place to round percentages in statistics.

If you want to calculate the percentage directly from the categorical data, here is another formula you may use.

$$\text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

Important Note

There are three ways to describe the proportion for categorical data: fraction, decimal, and percentage. Notice for the flu data example above, we have the three ways of describing the data: the fraction 41/326, the decimal proportion 0.126, and the percentage 12.6%. All of them are equivalent. It is important to be comfortable with fractions, decimal proportions and percentages when describing categorical data. They are a foundation for more advanced categorical analysis later on.

Calculating a Frequency (Count) from a Percentage

How to calculate a count (frequency) from a percentage or proportion. Sometimes a percentage is given in a scientific report or in an article. For more advanced proportion analysis, the computer programs usually require the actual count (frequency). So it is important to be able to find the frequency from percentage information.

Start by converting the percentage into a proportion.

$$\text{Proportion} = \text{Percentage} \div 100 \text{ (and remove the percent symbol \%)}.$$



Now multiply the proportion times the total to get the amount (frequency). This often called taking a “percentage of a total”. It is important to round your answer to the ones place since is the number of people or objects that have a certain characteristic.

Count (Frequency) = Decimal Proportion \times Total.

Example

According to the Center for Disease Control (CDC), about 32% of Americans have hypertension (high blood pressure). According to suburbanstats.org, Tulsa Oklahoma has approximately 603,403 people living in it. If the CDC is correct and 32% of Americans have hypertension, then how many people do we expect to have hypertension in Tulsa?

Step 1: Convert 32% into a decimal proportion.

$$32\% = 32 \div 100 = 0.32$$

Step 2: Multiply the decimal proportion by the total.

$$\text{Amount of people with hypertension} = 0.32 \times 603403 = 193088.96 \approx 193,089$$

So approximately 193 thousand people in Tulsa have high blood pressure. This is vital information for hospitals and doctors in the Tulsa, Oklahoma area.

Bar Charts and Pie Charts

A quick way to count how many people or objects have a certain label is to create a Bar Chart or Pie Chart. There are many different statistics software that we could use to create these graphs. They are useful to show the characteristics of categorical data.

Creating a Bar Chart with Raw Data and StatKey

StatKey does not create pie charts, but does have a nice bar chart feature. It not only creates the bar chart from the raw data but also calculates the counts (frequencies) from each category as well as the decimal proportions.

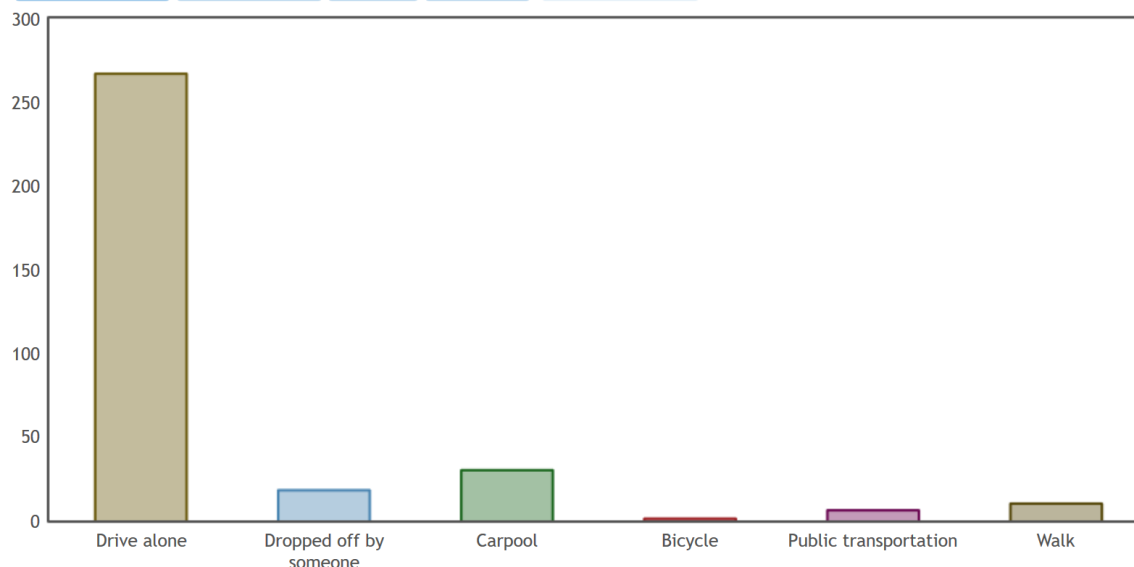
To make a bar chart with raw data, go to www.lock5stat.com and click on the “StatKey” button. Now click on “one categorical variable” under the descriptive statistics and graphs button. If you have raw categorical data, click the “edit data” tab and paste your raw categorical data into StatKey. Make sure to check “raw data” at the bottom. If your data has a title, also check “data has a header row”. No click “OK”.

For example, I copied and pasted the “transportation data” from the Math 140 Fall 2015 survey data at www.matt-teachout.org into StatKey and created the bar chart. Notice it not only created the graph, but also gave me the counts (frequencies) and the decimal proportions.



StatKey Descriptive Statistics for One Categorical Variable

Custom Dataset ▾ Show Data Table Edit Data Upload File Change Column(s)



Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

Creating a Bar Chart with Summary Data and StatKey

Categorical data is often summarized by the counts for each variable. When a data analyst receives categorical data to analyze, it may not be in raw form. Often it is just the counts (frequencies). In that case, when you go to the “edit data” button, you will need to type in the variables and counts as shown below. Uncheck the “raw data” box at the bottom and push “OK”. Note that you need only one space after the comma and do not type in the totals. Notice you will get the exact same graphs, counts and proportions as shown above.

Response, Frequency

Drive alone, 267

Dropped off by someone, 18

Carpool, 30

Bicycle, 1

Public Transportation, 6

Walk, 10



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Creating a Pie Chart with Raw Categorical Data and Statcato

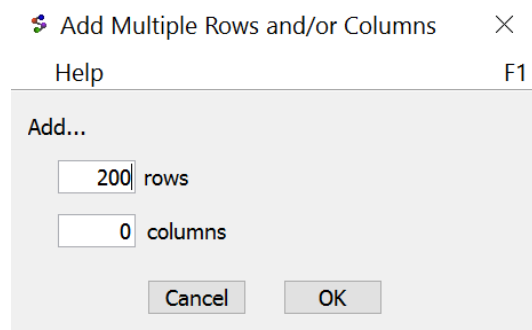
A pie chart is a very useful graph and can give the count (or frequency) for each variable and the percentages for each variable.

To create a pie chart with Statcato, open your excel spreadsheet. Copy and paste your column of categorical data from Excel into Statcato. Before pasting, be sure to click on the gray at the top of the column in Statcato, since titles must go in the gray. Now click on the graph menu at the top and then “pie chart”. Click on “data values from a worksheet” and then under “data” put in the column. If your data is in the first column, you will click on “C1”. If it is in the second column, you will click on “C2”, and so on. Give the chart a title and click on “Show Legends” and “Show Values/Percentages for each Pie Sector”. You can sort the graph by category or by frequency (counts). If you click on “sort by category”, the pieces will be put in alphabetical order clockwise around the circle. If you click on “sort by frequency,” then the chart will be organized from the smallest section to the largest section clockwise around the circle.

Graph Menu => Pie Chart => Data Values from a Worksheet => Sort by Categories or Frequencies, Show Legend, Show Values/Percentages

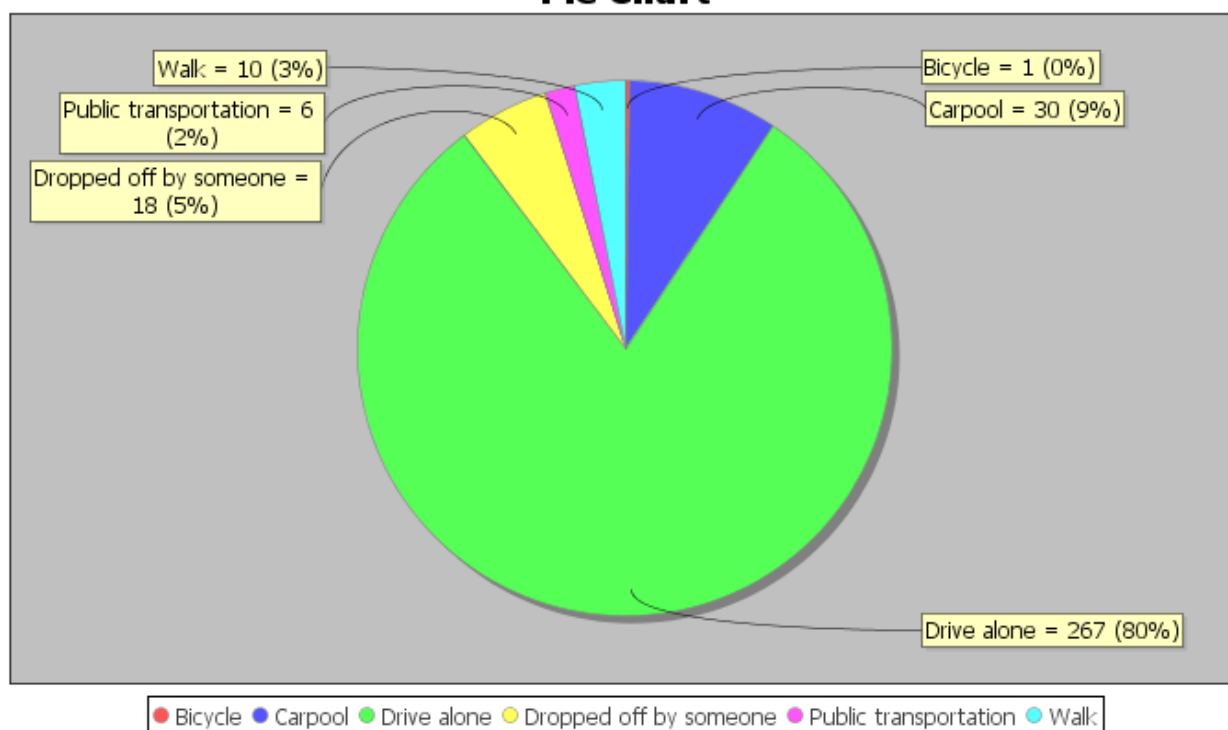
Let us use the same example, and open the transportation data from “Math 140 Survey Data” from fall 2015.

Important Reminder: If your data set is over 300 entries, you will need to add some rows to Statcato. The math 140-survey data had close to 350 students, so we will need to add some rows to the spreadsheet in Statcato before copy and pasting from Excel. (I added 200 more rows to Statcato before I tried to copy and paste.)



Once you have added enough rows in Statcato, copy and paste the column of data that says “Transportation” in Statcato. Do not forget to put the title in the gray cell at the top. Now go to the graph menu and make a pie chart. We will show two versions of the graph. One if you sort by categories and the other if you sort by frequencies. That way you can see the difference and which one you like better. The following graph was sorted by categories. Notice it gives the same counts as StatKey, though the proportions have been converted into percentages and rounded to two significant figures. You can copy and paste the graph into a Word or Pages document, by going to the “graph” button on the left side of the graph and click on “copy graph to clipboard”.



Pie Chart

Notice at the touch of a button, the computer can tell us all of the counts (frequencies) and all of the percentages. We can answer all sorts of questions about how these students get to the college.

Creating a Pie Chart or Bar Chart with Summary Data and Statcato

Categorical data is often given in summarized form with the variables and the counts. Statcato cannot make bar charts from raw data, but it can make a bar chart from summary counts. Statcato can also make a pie chart from summarized data. Suppose we do not have access to the raw categorical transportation data. Suppose we only knew the variable labels and the counts (frequencies) into two columns of Statcato. We will use the transportation data again. Note that titles like “variable” or “count” must be typed in the gray where it says “Var”.

	C1	C2
Var	Variable	Count
1	Drive Alone	267
2	Dropped off by someone	18
3	Carpool	30
4	Bicycle	1
5	Public Transportation	6
6	Walk	10

Now go to the graph menu and then “pie chart”. Click on “Summary Data from Worksheet”. Give the columns for the categories and the columns for the frequencies.



Pie Chart ×

Help F1

Graph Variables

☒ Summary Data from Worksheet

Categories: ▾

Frequencies: ▾

☐ Data Values from Worksheet

Data: ▾

Graph Options

Chart Title:

☒ Sort by Categories

☐ Sort by Frequencies

☒ Show Legends

☒ Show Values/Percentages for each Pie Sector

Notice the pie chart looks the same as the one we created with raw data.

We can also create a bar chart from summary categorical data. Again, type in the summary counts and variables into two columns of Statcato. Then go to the graph menu in Statcato and click on “Bar Chart”. Statcato will want to know what column has your variable names and the column that has your counts.

Under “Select the column variable of a new series”, pick the column with your counts (frequencies). Mine was in column 2. Now click “Add Series”. Under “Select the column variable containing categories” select the column that has your variable names. Mine was in column 1. Type in a title and “show legend” and press OK. You can make the bars vertical or horizontal as well. I used vertical in this example.

	C1	C2
Var	Variable	Count
1	Drive Alone	267
2	Dropped off by someone	18
3	Carpool	30
4	Bicycle	1
5	Public Transportaion	6
6	Walk	10



Bar Chart

Help F1

Graph Variables

Graph Series

C2 Count

Select the column variable of a new series:

C2 Count

Add Series

Select the series to be removed:

Remove Series

Categories

Select the column variable containing categories:

C1 Vari...

Direction of Bars

☐ Horizontal

☒ Vertical

Graph Options

X-axis Label:

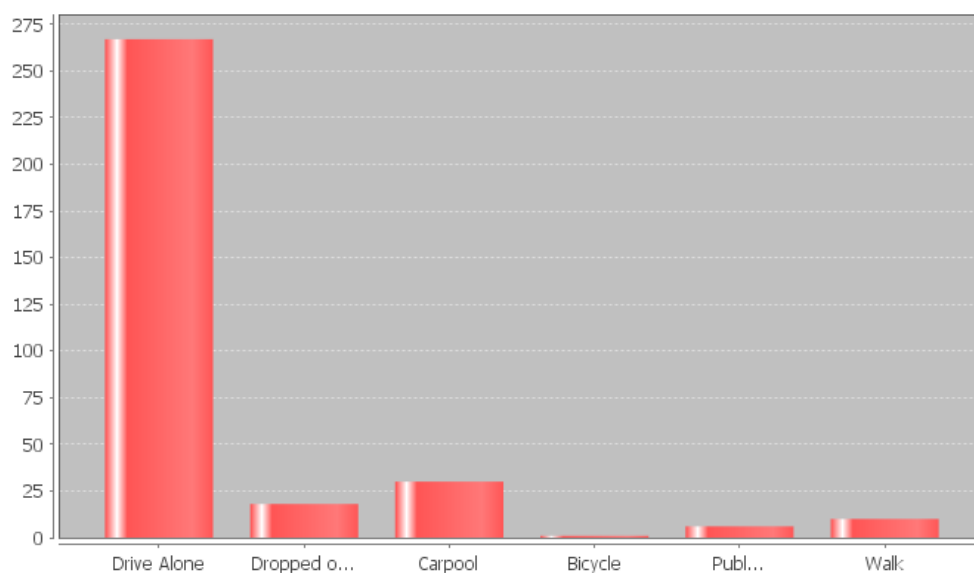
Y-axis Label:

Plot Title: Bar Chart

☒ Show Legend

OK Cancel

Transportation Bar Chart



Comparing Percentages

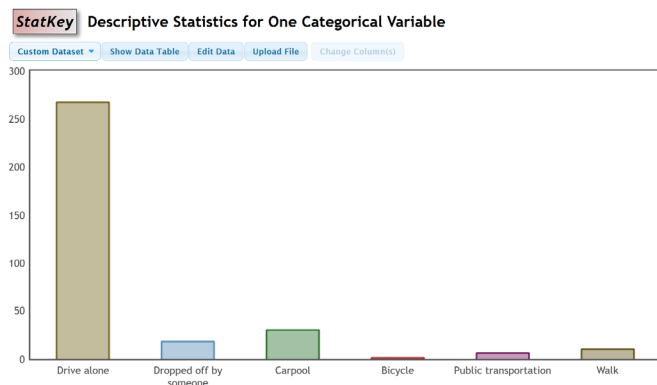
Sometimes we want to compare categorical variables and see if one variable has a significantly higher proportion or percentage than another. To compare proportion or percentages, many people often calculate the “percentage of increase”. There are three different ways of calculating the percentage of increase. Any of these formulas give the same answer.

$$\text{Percent of Increase} = \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%$$

$$\text{Percent of Increase} = \frac{(\text{Higher \%} - \text{Lower \%})}{\text{Lower \%}} \times 100\%$$



For example, let us look at the transportation bar chart found with StatKey. Suppose we want to compare the percentage of math 140 students that carpool verse the percentage that were dropped off. We can calculate the percent of increase from the counts, proportions or percentages. It is important to recognize which is the lower count (frequency) and which is the higher count. In this case, the number of students that carpool was higher than the number of students that were dropped off. The key question is was it significantly higher.



Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

We can calculate the percent of increase from either the proportions or the percentages.

$$\text{Percent of Increase} = \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\% = \frac{(0.09 - 0.054)}{0.054} \times 100\% \approx 66.7\%$$

$$\text{Percent of Increase} = \frac{(\text{Higher \%} - \text{Lower \%})}{\text{Lower \%}} \times 100\% = \frac{(9\% - 5.4\%)}{5.4\%} \times 100\% \approx 66.7\%$$

Notice this tells us that the proportion of students that carpool is 66.7% higher than the proportion that are dropped off. This difference seems statistically significant.

Note: In chapter 3 and chapter 4, we will learn how to use confidence intervals, test statistics, and P-values to determine significant differences. These are generally more accurate than the percent of increase calculation.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Statistical Significance versus Practical Significance

Sometimes when there is a statistically significant difference, it does not necessarily mean it is of practical use. In the last example, we saw that the number of students that carpool was a 66.7% higher than the number of students that are dropped off. Does this mean that college should make a special parking lot for all of the Math 140 students that carpool? Probably not. We are only talking about a difference of 12 total students a semester. College of the Canyons has thousands of students. So even though the percent of increase is significant, the data is not really of practical use in the sense that I would be careful of making huge decisions from the 66.7%.

Binomial Proportions with Statcato (Optional Topic)

Sometimes we want to know a percentage or proportion associated with a categorical event happening multiple times. One example of this is called a binomial proportion. A binomial proportion can be calculated from categorical data with only two outcomes (winning or losing, smoking or not, drinking alcohol or not). These are often referred to as “success” and “failure”. The individuals must be independent of each other and the event (success) percentage (p) must be the same all the time. To calculate a binomial percentage, you will need a computer program and three bits of information, the number of events (number of successes), the event proportion (p), and the sample size (n).

Example

Categorical data often has a requirement of at least 10 success and at least 10 failures. Suppose we collect a random sample of 72 people and ask them whether they smoke cigarettes or not. Is 72 a large enough data set? Are we likely to get 10 or more people that smoke and 10 or more people that do not smoke? We can use Statcato to calculate this binomial percentage. According to the center for disease control, about 15.5% of adults in the U.S. smoke cigarettes.

Probability (percentage) of 10 or more people smoking =?

Number of Trials = Sample Size (n) = 72

Number of Events (X) = 10

Event Probability (p) = 0.155

Calculating binomial percentages can be challenging. Here is the formula that computer programs use.

Binomial Probability of X events: $P(X) = C(n, x)p^x(1 - p)^{n-x}$

The problem with this formula is we have to calculate it for $X = 10, X = 11, X = 12, \dots, X = 72$ and then add all the proportions together. That is very difficult. It is best to let a computer program do the heavy lifting.

Open Statcato and click on “Calculate” menu. Then click on “probability distributions” and “binomial”. Statcato is limited in the sense that it only calculates binomial percentage for either equal to (probability density) or less than or equal to (cumulative probability). So if we are calculating a greater than question, we must think about the opposite (less than or equal to). In this problem, we want to find 10 or more. The opposite of this would be 9 or less. Therefore, we will calculate the percentage for 9 or less, and then subtract the answer from 100%. This is sometimes called a “complement” proportion. In Statcato, put in the following. Under “Number of trials”, put in the sample size 72. Under “constant” put in the number of events 9. Under “Event probability”, put in 0.155. Now push the “Cumulative Probability” button and push “compute”.



Binomial Probability Distribution

Help F1

Distribution

Distribution Parameters:

Number of trials: 72

Event probability: 0.155

Compute:

☐ Probability density

☒ Cumulative probability

☐ Inverse cumulative probability

Inputs and Outputs

Input(s):

☐ Column: v

☒ Constant: 9

Store Results in: (optional)

(e.g. C1 for column label, or variable name)

Compute Close

Binomial Distribution: $n=72$, $p=0.155$

Input: 9.0

Type: Cumulative probability

X P(\leq X)

9.0 0.304036

Notice the probability of getting 9 or less is 0.304 or 30.4%. This is the complement percentage to what we are looking for. So the probability of getting 10 or more people that smoke should be $100\% - 30.4\% = 69.6\%$. This may not be a high enough percentage to assure us that we will get at least 10 people that smoke. I would recommend collecting more data (increase the sample size).

Example

Suppose a person is playing a game of roulette that has a $1/38$ or 2.63% chance of winning. The gambler plans to play the game 20 times. What is the probability that he or she wins just once?

Open Statcato and click on "Calculate" menu. Then click on "probability distributions" and "binomial". Remember to calculate equal, you need to click on the "probability density button".

Number of Trials = 20

Event Probability = 0.0263

Number of Events = 1 (Put this in the "constant" box.)



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Binomial Probability Distribution
×

Help
F1

Distribution

Distribution Parameters:

Number of trials:

Event probability:

Compute:

☒ Probability density
☐ Cumulative probability
☐ Inverse cumulative probability

Inputs and Outputs

Input(s):

☐ Column:
☒ Constant:

Store Results in: (optional)

(e.g. C1 for column label, or variable name)

Compute
Close

Binomial Distribution: $n=20$, $p=0.0263$

Input: 1.0

Type: Probability density

X P(X)

1.0 0.317003

Notice the answer can be found under "P(X)". So the gambler has a 0.317 (31.7%) chance of winning the game once.



Problem Set Section 1E

1. Convert each of the following percentages into a proportion. Do not round the answers.

- a) 75%
- b) 2.75%
- c) 0.664%
- d) 0.082%
- e) 39.7%
- f) 8.6%
- g) 0.189%
- h) 0.0025%
- i) 3.16%
- j) 250%
- k) 96.1%
- l) 0.48%
- m) 0.007%
- n) 8.73%
- o) 66.2%
- p) 9%
- q) 100%

2. Convert each of the following proportions into a percentage. Do not round the answers.

- a) 0.057
- b) 0.812
- c) 0.0033
- d) 0.0214
- e) 0.0613
- f) 0.451
- g) 0.00045
- h) 0.0779
- i) 0.046
- j) 0.3161
- k) 0.0027
- l) 0.051
- m) 0.0058
- n) 0.847
- o) 1
- p) 0.00022
- q) 0.0204

(#3-10) *Directions: Convert the given percentages into proportions. Then use the following formula to find the estimated amounts. Round your answers to the ones place.*

$$\text{Estimated Amount} = \text{Proportion} \times \text{Total}$$

3. According to an article by CBS news, approximately 15% of Americans still do not have health insurance. If approximately 78,300 people live in Chino Hills CA, then how many people in Chino Hills would we expect to not have health insurance? Round your answer to the ones place.

4. According to an article online, about 30% of Americans own at least one gun. About 305,700 people live in Stockton CA. If the article was accurate, then approximately how many people in Stockton do we expect to own at least one gun? Round your answer to the ones place.

5. An article by the American Diabetes Association estimates that as of 2012, about 9.3% of Americans have diabetes. College of the Canyons has approximately 18,400 students. If the percentage were correct, how many COC students would we expect to have diabetes? Round your answer to the ones place.



*This chapter is from Introduction to Statistics for Community College Students,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

6. According to a news report by www.nielsen.com, about 15.9% of Americans struggle with hunger. Lancaster CA has approximately 161,000 people living in it. If the percentage from the Nielsen report is accurate, then how many people in Lancaster CA may be struggling with hunger? Round your answer to the ones place.

7. According to an article by the Autism Society, about 1.47% of people in the U.S. have autism. The article also stated that the percentage is increasing every year and that Autism is one of the fastest growing disorders in the U.S. Van Nuys, CA has approximately 136,400 people living in it. If the percentage by the Autism Society is correct, how many do we expect to have autism?

8. According to a recent article, about 0.51% of airbags in the U.S. are defective. According to vehicle registration data, there are approximately 1,769,000 cars in San Francisco, CA. How many of them do we expect to have defective airbags?

9. According to a recent U.S. census, about 14.8% of people in the U.S. live below the poverty line. About 305,700 people live in Stockton CA. If the census was accurate, then approximately how many people in Stockton are living in poverty?

10. According to an article by the American Medical Association, approximately 33% of medical doctors in the U.S. have been sued by patients for malpractice. Suppose a hospital has currently 147 doctors on staff. How many of them do we expect to have been sued for malpractice?

(#11-15) *Directions: Use the following formulas to calculate the proportions, percentages and the percent of increase. Then answer the given questions.*

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}}$$

$$\text{Percentage} = \text{Decimal Proportion} \times 100\%$$

$$\text{Percent of Increase} = \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%$$

11. An article at www.seattletimes.com was addressing the issue of whether women in the U.S. prefer traditional jeans or athletic wear like yoga pants, sweat pants or leggings. Assume that a random sample of 213 total women were asked if they prefer traditional jeans or athletic wear. Assume 139 said they prefer athletic wear and 74 said they prefer traditional jeans. Calculate the decimal proportions and the percentages for both athletic wear and traditional jeans. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

12. The article at www.seattletimes.com also said that jean companies are creating more and more stretchy jeans to compete with the growing trend of women preferring athletic wear. Assume that a random sample of 197 total women were asked if they prefer stretchy jeans or athletic wear. Assume 103 said they prefer athletic wear and 94 said they prefer stretchy jeans. Calculate the decimal proportions and the percentages for both athletic wear and stretchy jeans. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

13. A hospital is trying to decide how to allocate resources to various departments. In particular, they are comparing the medical/surgical ward to the telemetry (heart monitor) ward since these wards have similar costs per patient. Assume we looked at a random sample of patients admitted to the hospital. Of the 350 total patients, 57 were admitted to the medical/surgical ward and 49 were admitted to telemetry. Calculate the decimal proportions and the percentages for both medical/surgical and telemetry. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

14. A company found that of their 348 total employees, 96 employees have health insurance and 252 employees do not have health insurance. Calculate the decimal proportions and the percentages for both having health insurance and not having health insurance. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



15. An experiment was done to test the effectiveness of a new medicine to treat depression. They found that of the 57 people that received the medicine, 13 indicated significant improvement in their depression symptoms. Of the 61 people in the placebo group, 11 indicated significant improvement in their depression symptoms. Calculate the decimal proportions and the percentages for the medicine and placebo groups. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

(#16-20) Directions: Go to www.matt-teachout.org, click on the “statistics” tab and then “data sets”. Open the indicated data set and copy the indicated column of categorical data. Go to www.lock5stat.com and click on StatKey. Under the “descriptive statistics and graphs” menu, click on “one categorical variable”. Click on the “edit data” button and paste in the column. Check the box for “raw data” and “data has a header row” and push OK. Then answer the questions. Use the following formula for the percent of increase calculation.

$$\text{Percent of Increase} = \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%$$

16. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the campus data. Use StatKey to make a bar chart, and a summary of the proportions and counts. What proportion of the students went to Valencia? What proportion of the students went to the Canyon Country campus? Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

17. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the gender data. Use StatKey to make a bar chart, and a summary of the proportions and counts. What proportion of the students identified as female? What proportion of the students identified as male? Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

18. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the hair color data. Use StatKey to make a bar chart, and a summary of the proportions and counts. Which hair color had the highest proportion? Which hair color had the lowest proportion?

19. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the political part data. Use StatKey to make a bar chart, and a summary of the proportions and counts. What proportion of the students identified as democratic? What proportion of the students identified as republican. Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

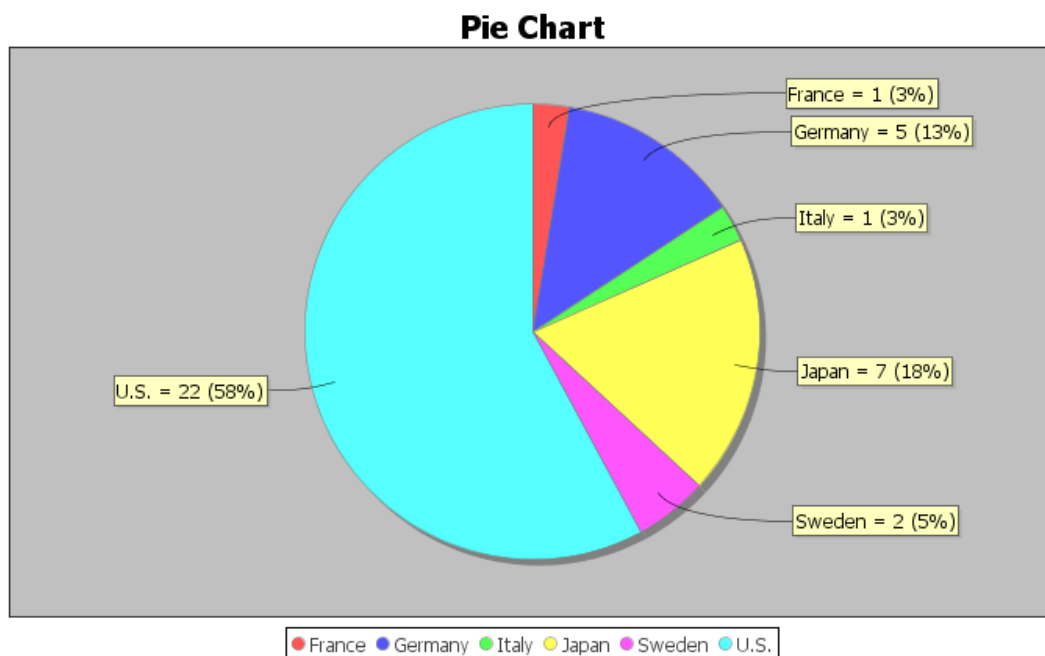
20. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the “month of birthday” data. This data has numbers in it. Explain why this is categorical data and not quantitative. Use StatKey to make a bar chart, and a summary of the proportions and counts. Which month had the highest percentage? Which month had the lowest percentage?

(#21-25) Use the following pie charts from Statcato to answer the following questions. Use the following formula for the percent of increase calculation.

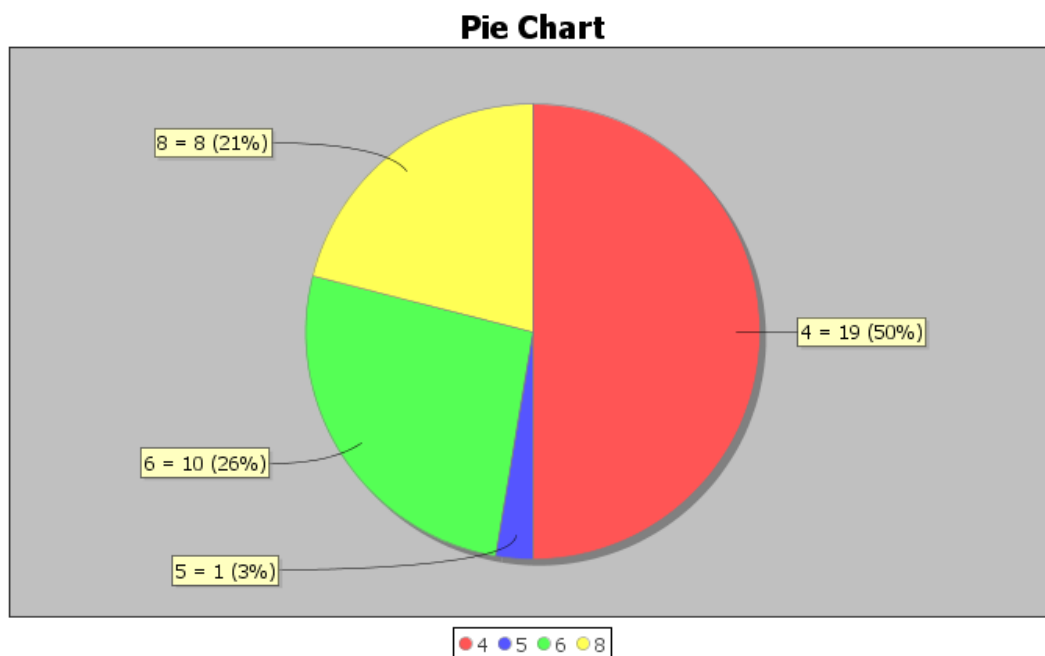
$$\text{Percent of Increase} = \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%$$



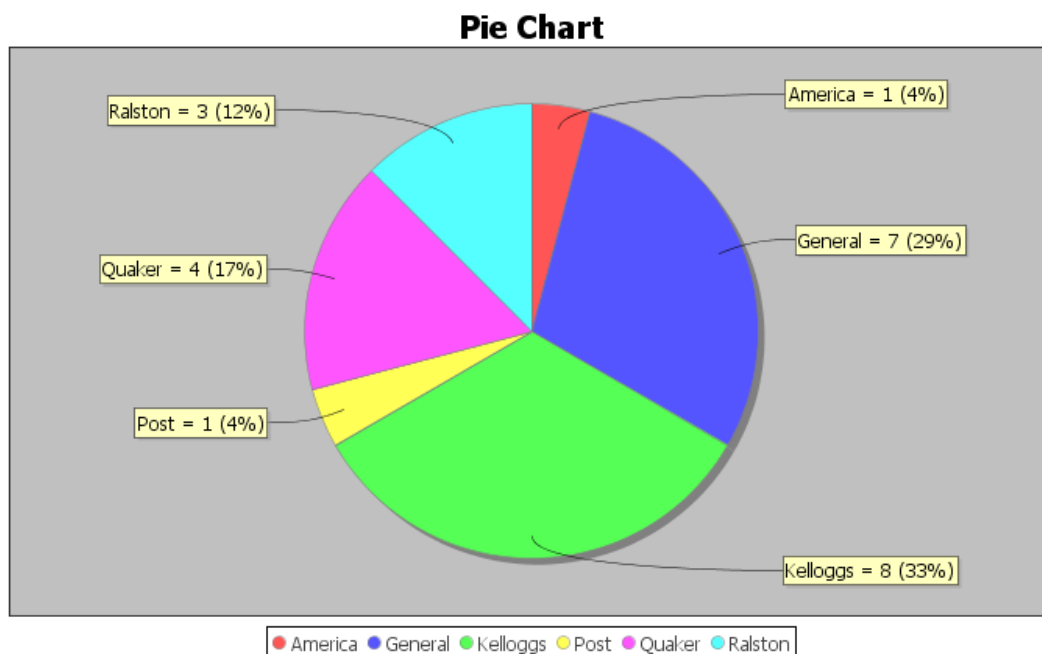
21. The following pie chart was created from the “car data” at www.matt-teachout.org. What percentage of the cars were made in France? How many of the cars were made in the U.S.? What proportion of the cars were made in Sweden? Calculate the percent of increase to compare Japan and Germany. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



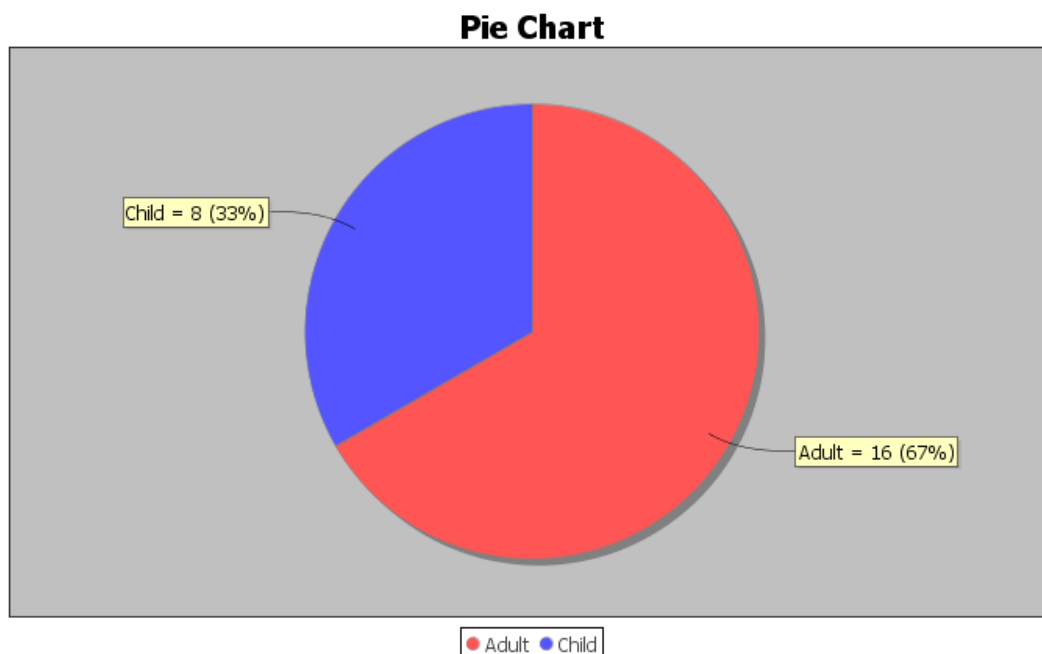
22. The following pie chart was created from the “car data” at www.matt-teachout.org. What percentage of the cars four cylinders? How many of the cars have eight cylinders? What proportion of the cars six cylinders? Calculate the percent of increase to compare four and eight cylinder cars. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



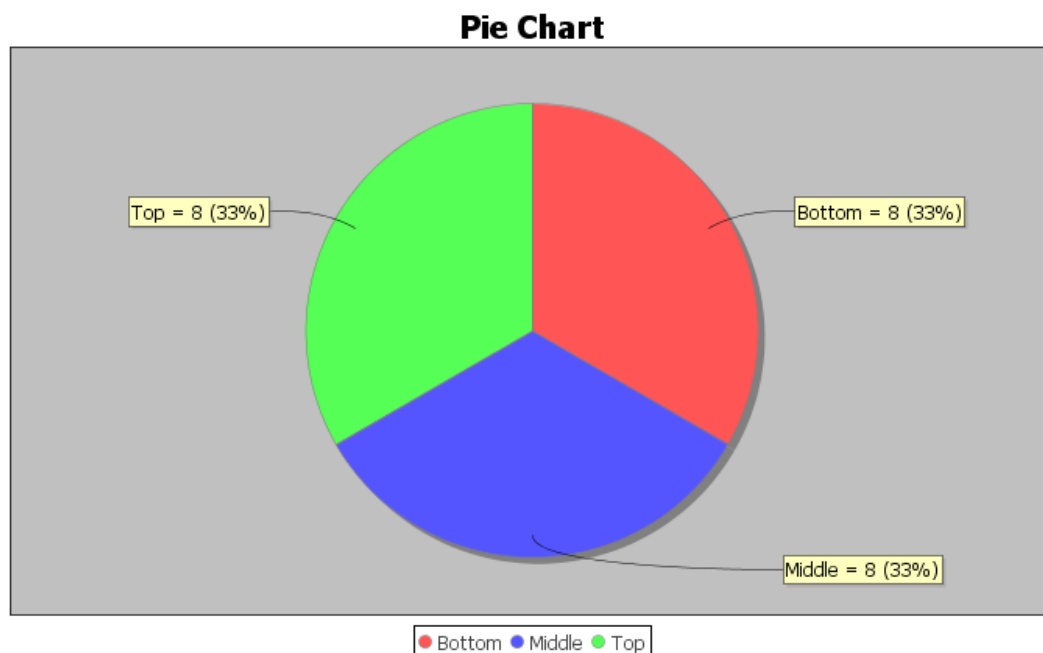
23. The following pie chart was created from the “cereal data” at www.matt-teachout.org. What percentage of the cereals did Quaker make? How many of the cereals did Ralston make? What proportion of the cereals did General make? Calculate the percent of increase to compare Kelloggs and Quaker. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



24. The following pie chart was created from the “cereal data” at www.matt-teachout.org. What percentage of the cereals were targeted toward adults? What percentage of the cereals were targeted toward children? Calculate the percent of increase to compare adult cereals and children cereals. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



25. The following pie chart was created from the “cereal data” at www.matt-teachout.org. What percentage of the cereals are displayed on the top shelf? How many of the cereals are displayed on the bottom shelf? What proportion of the cereals are displayed on the middle shelf? Calculate the percent of increase to compare the top and bottom shelf cereals. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



Optional Binomial Probability Questions

(#26-30) *Directions: Open Statcato on either one of the in-class or TLC computers. Go to the “calculate” menu, click on “probability distributions” and then “binomial”. Enter the total under “number of trials” and the proportion under “event probability”. Under “constant” put in the number of successes. Check “probability density” if you want to calculate equal to. Check “cumulative probability” to calculate less than or equal to. For greater than or equal to, subtract your less than or equal to (opposite) answer from one. Assume the questions meet the requirements for calculating a binomial probability.*

26. To win at a dice game, the player must role two dice and get a 7 or 11 sum. This game has a 22.2% chance of winning. Suppose a player rolls the dice 18 times.

- What is the probability that they win exactly once?
- What is the probability that they win two times or less?
- What is the probability that they do not win at all? (This means she wins zero times.)
- What is the probability that they win three times or less?
- What is the probability that they win four or more times? (Subtract your answer in (d) from one.)
- What is the probability that they win four times or less?
- What is the probability that they win five or more times? (Subtract your answer in (f) from one.)



27. A car company thinks that their minivan transmissions have a 12% defective rate. A total of 84 minivans were brought in to a service center this month.

- a) What is the probability that exactly 11 of them need to have their transmission replaced?
- b) What is the probability that exactly 8 of them need to have their transmission replaced?
- c) What is the probability that 12 or less of the minivans will need their transmission replaced?
- d) What is the probability that 13 or more of the minivans will need their transmission replaced?
(Subtract your answer in (c) from one.)
- e) What is the probability that 6 or less of the minivans will need their transmission replaced?
- f) What is the probability that 7 or more of the minivans will need their transmission replaced?
(Subtract your answer in (e) from one.)

28. Suppose we take a random sample of 57 total people and ask them if they smoke cigarettes or not. Assume that the population percentage for smoking in the U.S. is 15.5%.

- a) What is the probability that we will get 9 or less people that smoke in the data set?
- b) We need to have at least 10 people in the data set that smoke. What is the probability that we will get 10 or more people that smoke in the data set? (Subtract your answer in part (a) from 1.) Is this percentage high enough for us to be confident that 57 people is a large enough data set? Explain.

29. Suppose we take a random sample of 57 total people and ask them if they smoke cigarettes or not. Assume that the population percentage for non-smokers in the U.S. is 84.5%.

- a) What is the probability that we will get 9 or less people that do not smoke in the data set?
- b) We need to have at least 10 people in the data set that do not smoke. What is the probability that we will get 10 or more people that smoke in the data set? (Subtract your answer in part (a) from 1.)

30. Suppose a person is playing a game of roulette that has a 2.63% probability of winning. The person plays the game forty times.

- a) What is the probability that they do not win at all? (The probability they win zero times.)
 - b) What is the probability that they win exactly one time?
 - c) What is the probability that they win two or less times?
 - d) What is the probability that they win three or more times? (Subtract your answer in (c) from one.)
 - e) What is the probability that they win one or less times?
 - f) What is the probability that they win two or more times? (Subtract your answer in (e) from one.)
-



Section 1F – Normal Quantitative Data Analysis

Vocabulary

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Normal Data: Data that is bell shaped, symmetric and unimodal. Also referred to as data that has a normal distribution.

Sample Size: Also called the total frequency.

Average: Also called the center of the data. A single number that represents a typical person or object in the data set.

Variability: Also called the spread. A measure of how spread out a data set is. A large spread tells us that the data is less consistent and the more difficult to predict. A small spread tells us that the data is more consistent and easier to predict.

Mean Average (\bar{x}): The balancing point for distances in a data set. The average for a data set that is normal.

Standard Deviation: The average or typical distance that points in a data set are from the mean. The measure of typical spread (typical variability) for a data set that is normal.

Maximum: The largest number in a data set.

Minimum: The smallest number in a data set.

Outliers: Unusual values in the data set.

Introduction

When analyzing numerical quantitative data, always start with finding the shape of the data set. Categorical data can be graphed, but does not have a shape. Categorical bar charts can be organized in a variety of ways depending on the order of the categories. Quantitative data is numerical measurement data and does have a shape.

Why should we find the shape?

The goal in analyzing quantitative data is to find the average, spread and unusual values. In statistics, there are many types of averages, many types of spreads. Shape helps us determine which averages and spreads are most accurate for the data.

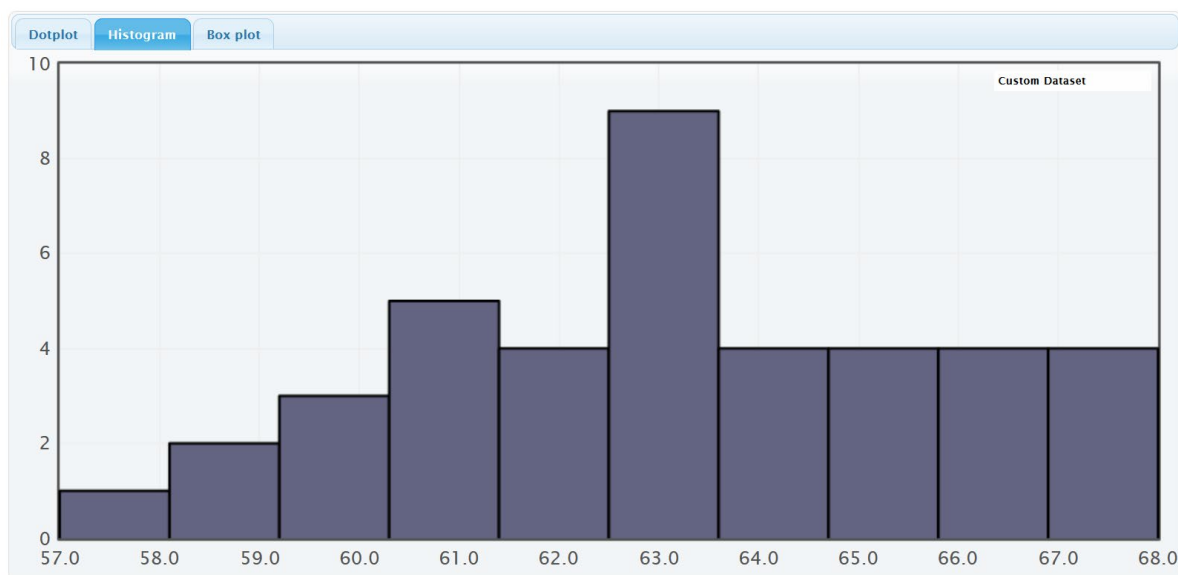
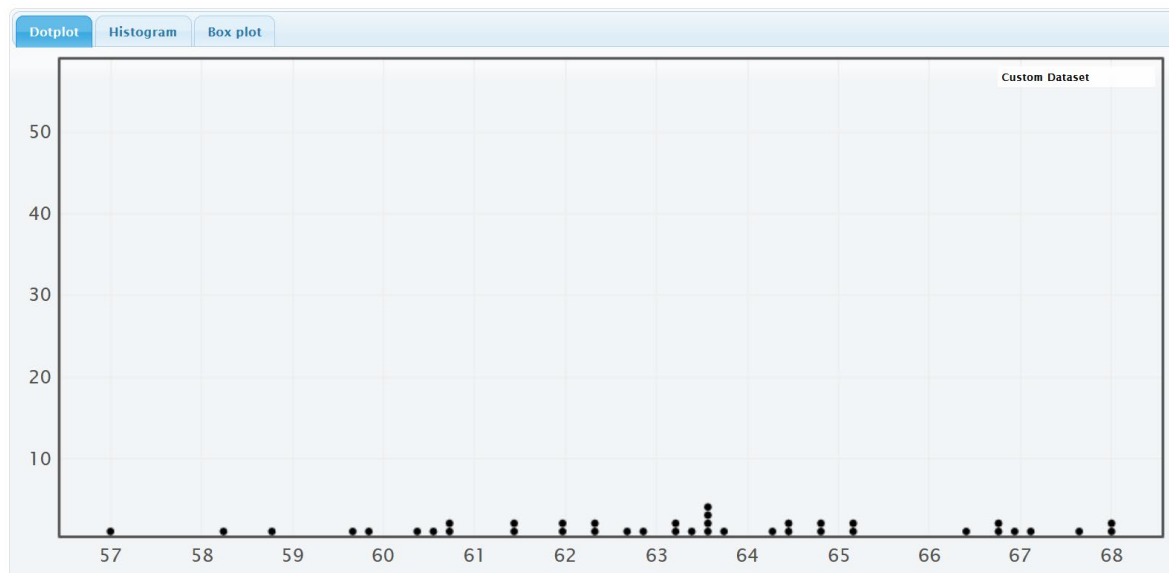
Quantitative Statistics and Graphs with StatKey

The most common quantitative statistics we like to look at are the mean, median, standard deviation, first quartile, third quartile, interquartile range, max, min, and range. The most basic kind of graph for quantitative data is the dot plot. The computer draws the numerical scale usually horizontally. It then draws a dot for every single number in the data set. Another type of graph is a histogram. This graph counts the number of data values in certain sections and makes a bar telling us how many numbers are in that section. The number of bars are also called “bins” or “buckets”. Another graph we like to look at is the boxplot. A boxplot is a graph of the first quartile, median, and third quartile as well as potential outliers.

All of these graphs and statistics can be made with StatKey. Let us look at an example. Go to www.matt-teachout.org and click on the “statistics” tab and then the “data sets” tab. Look for the “Health Data” excel file. Open the data set and copy the women’s heights data. Notice the data is quantitative. It measures the height in inches of the women and it seems reasonable to look for an average height of these women.



Go to www.lock5stat.com and click on the “StatKey” button. Under the “Descriptive Statistics and Graphs” menu, click on “One Quantitative Variable”. Click on the “Edit Data” button. Copy and paste the women’s height data into StatKey. Uncheck the box that says, “First column is an identifier”. An identifier is a word next to every number. This data set does not have that. Check the box that says, “Data has a header row”. This means the data set has a title. Now push “OK”. Notice StatKey gives you the sample statistics, a dot plot, a histogram and a boxplot.



On the right of this histogram, you will see a slider that can adjust the number of “buckets” or bins. The smaller the data set the less bins you should have. This data set only has 40 numbers, so we want only a few bars. If we slide it to 3 buckets, we get the following.

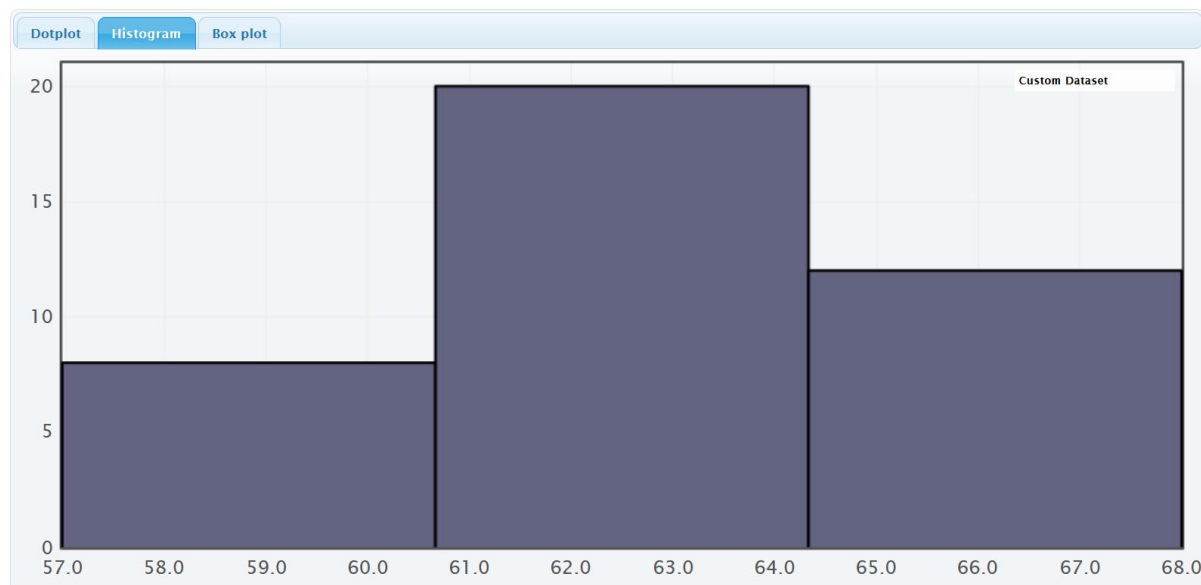
Histogram Controls

[Set Limits](#)

Number of buckets: 3



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18



This data has a very special shape. It is called bell shaped or normal. Normally distributed data has the highest bar in the middle and about equal number of bars decreasing from the middle. It looks like bell. We see that this data set is relatively normal (bell shaped) or “normally distributed”. StatKey has also given us summary statistics. Which statistics are most accurate for normal data?

Summary Statistics

Statistic	Value
Sample Size	40
Mean	63.195
Standard Deviation	2.741
Minimum	57
Q ₁	61.350
Median	63.350
Q ₃	64.900
Maximum	68

Mean and Standard Deviation

Important Note about Shape: The mean and standard deviation should only be used if the data set is normal. The mean and standard deviations are not accurate if the data does not have a normal shape.

Mean (\bar{x}): The mean is a type of average used for data that is normally distributed. The mean balances the distances between all the numbers in the data set and the mean. Think of it this way. If you took all the numbers in the data set below the mean, measured their distances from the mean, then added up those distances. That total distance for numbers below the mean would be equal to the total distance for numbers above the mean. The mean is calculated by adding up all the numbers in a data set ($\sum x$) and then dividing by how many numbers are in the data set (sample size “n”).

$$\bar{x} = \frac{\sum x}{n}$$



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Standard Deviation (S): We said that the mean balances the distances in a data set. The standard deviation calculates the average distance numbers are from the mean. It is the most accurate measure of typical spread for data sets that are normally distributed. To calculate the standard deviation, computer programs take every single number in the data set and subtract the mean. Since those differences can be negative sometimes, they computer squares all the differences and then adds up the squares. This is a famous calculation called “sum of squares”. Since we want the average distance, we divide by $n - 1$ (degrees of freedom) and take the square root at the end to undo all the squares. Never calculate this by hand. It is a long calculation that should be left to a computer program.

$$S = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Why do we study spread? Spread is a measure of how much variability is in the data set. Think of it this way. Suppose we were looking at exam scores in a history class that are normally distributed. If a data set were very spread out, then the standard deviation would be quite large. This would mean that the scores had a lot of variability. We had A's, B's, C's, D's, and F's. The exam scores are not consistent, and the history teacher will have a hard time predicting how her class will do. If the data set has a small spread, then the standard deviation would be quite small. The exam scores are very consistent. Maybe everyone in the class got an A or a high B. It is easier to predict how the class will do.

Statistics for Normal Data

Quantitative Variable and Units

Sample Size (n)

Maximum Value

Minimum Value

Average: Mean (\bar{x})

Spread: Standard Deviation (s)

Typical Values: One standard deviation from the mean. Here is a formula that is sometimes used.

$$\bar{x} - s \leq \text{typical values} \leq \bar{x} + s$$

Outliers (unusual values): More than two standard deviations from the mean. Here are formulas that are sometimes used.

$$\text{Unusually Low Values (Low outliers)} \leq \bar{x} - 2s$$

$$\text{Unusually High Values (High outliers)} \geq \bar{x} + 2s$$

Women's Height Example

Quantitative Variable and Units: Women's heights in inches

Sample Size (n): There were 40 women in the data set.

Maximum Value: The tallest woman in the data set was 68 inches.

Minimum Value: The shortest woman in the data set was 57 inches.

Average: Mean (\bar{x}). The average height of the women in the data was 63.195 inches.

Spread: Standard Deviation (s). The typical spread for this data was 2.741 inches. Typical women in the data were 2.741 inches from the mean.



Typical Values: Add and subtract the mean and standard deviation. Typical women in the data set have a height between 60.454 inches and 65.936 inches. We will see later that these values are the cutoffs for the middle 68% for normal data.

$$\bar{x} - s \leq \text{typical values} \leq \bar{x} + s$$

$$63.195 - 2.741 \leq \text{typical values} \leq 63.195 + 2.741$$

$$60.454 \leq \text{typical values} \leq 65.936$$

Outliers (unusual values): Add and subtract the mean and two standard deviations. Unusually tall women are 68.677 inches or higher. There are no unusually tall women in this data set. Unusually short women are 57.713 inches or lower. This means that the minimum value of 57 inches was unusually low. We will see later that these values are the cutoffs for the top and bottom 2.5% for normal data.

$$\text{Unusually Low Values (Low outliers)} \leq \bar{x} - 2s = 63.195 - (2 \times 2.741) = 57.713 \text{ inches}$$

$$\text{Unusually High Values (High outliers)} \geq \bar{x} + 2s = 63.195 + (2 \times 2.741) = 68.677 \text{ inches}$$

Quantitative Statistics and Graphs with Statcato

You can also make dot plots, histograms and sample statistics with Statcato. Copy and paste women's heights into a column of Statcato. The data set is only 40 values, so you will not need to add rows to Statcato. To make a dot plot, go to the graph menu and click on dot plot. Then click on the column of data you want to use. Then push ok.

Making a dot plot in Statcato: *Graph => Dot plot => Pick a column => OK*

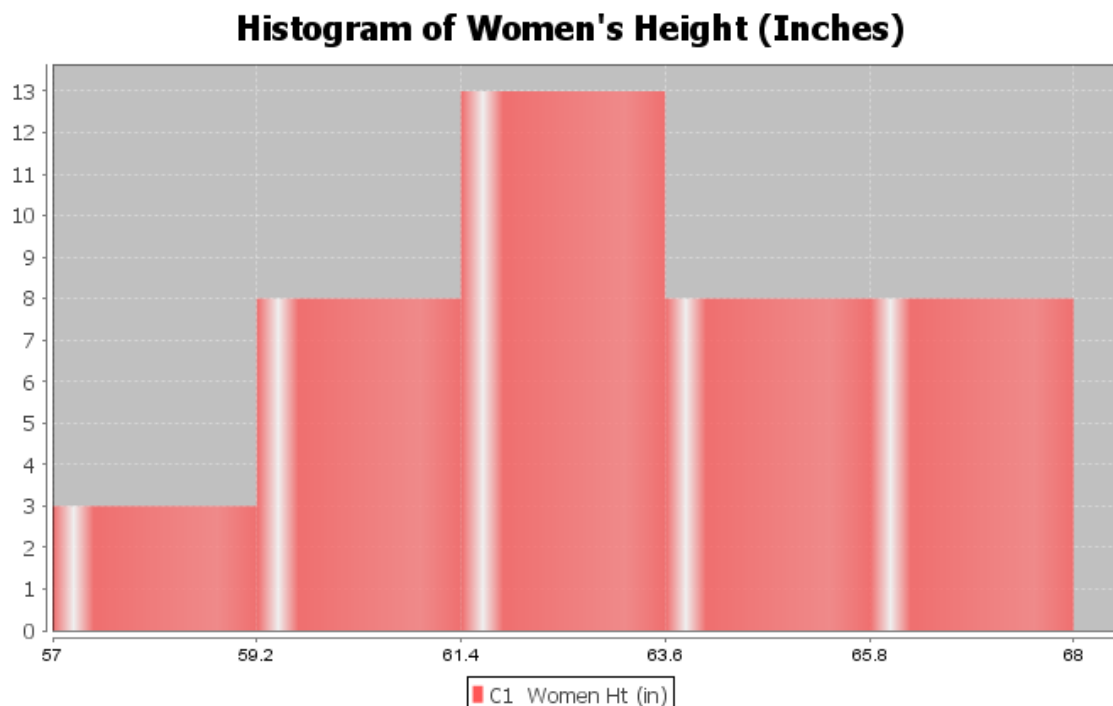
Here is the dot plot for the 40 women's heights.



To make a histogram in Statcato, go to the graph menu, and then click on histogram. Chose a column of data and how many bars (bins) you want. Then chose ok.

Making a histogram in Statcato: *Graph => Histogram => Pick a column => Chose number of bins => OK*

Note about bins: *If you chose too many bars then the histogram starts to look very crazy and you will have a hard time seeing the shape. Remember the goal is to break the dots up into groups. For example, in this health data there are only 40 women. I would not want 40 bins since that would give me about one bar per dot. If it were a small data set like the health data, I would do about three bins. Remember, the more bins you have, the more difficult it is to see the shape. This graph has five bins.*



Notice again that the highest bar is close to the middle and the bars get smaller as we move away from the middle. This is often called “Bell Shaped” or “Normal Data”. Some like to describe this shape as unimodal (1 hill) and symmetric (left and right side look about the same). I prefer to call it bell shaped or normal.

We can also calculate all of the sample statistics with Statcato. Go to the “Statistics” menu. Then click “Basic Statistics” and “Descriptive Statistics”. I had pasted the data into column 1, so type in “C1” under “input variable”. Check the boxes for statistics that you want and push “OK”.



Descriptive Statistics

Help

Inputs

Input Variable(s):
C1

Enter valid column names separated by space. For a continuous range of columns, separate using dash (e.g. C1-C30).

By Variable (optional):

Results

Store Results in:
☐ New datasheet

Statistics

☐ Select all statistics

☒ Mean
☐ SE of mean
☒ Standard deviation
☐ Variance
☐ Coefficient of variation

☐ Trimmed mean: cutoff % % of values to be trimmed (between 0 and 100)
☐ Sum
☒ Minimum
☒ Maximum
☐ Range

☒ First quartile
☒ Median
☒ Third quartile
☒ Interquartile range
☐ Mode
☐ Percentile:
e.g. 10 for the 10th percentile

☐ N nonmissing
☐ N missing
☒ N total
☐ Cumulative N
☐ Percent
☐ Cumulative Percent

☐ Sum of squares
☐ Skewness
☐ Kurtosis
☐ MSSD

OK Cancel

Z-scores

In normal data, we often want to find out how many standard deviations a number (X-value) is from the mean. This is called a “Z-score”. Here is a common formula. In later chapters, we will see that we can also use the Z-score as a test statistic to measure significance.

$$Z = \frac{(X \text{ value} - \text{Mean})}{\text{Standard Deviation}}$$

Example: In the last example, we saw that the women’s height data was normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. Suppose a woman is 72 inches tall. What would be the Z-score for her height? Is she unusually tall?

It is important when calculating a Z-score that you subtract the X value and the mean first. Then divide by the standard deviation. Most people in statistics round Z-scores to the hundredths place (two numbers to the right of the decimal).

$$Z = \frac{(X \text{ value} - \text{Mean})}{\text{Standard Deviation}} = \frac{(72 - 63.195)}{2.741} = +3.21233 \approx +3.21$$

If the X-value is below the mean, the Z-score will be negative. If the X-value is above the mean, the Z-score will be positive. This Z-score was positive. So the woman that is 72 inches tall is 3.21 standard deviations above the mean. Is this unusual?

Remember the formula above for finding the cutoff for unusual values for normal data. Notice it is two standard deviations above and below the mean. Two standard deviations above the mean would be a Z-score of +2. Two standard deviations below the mean would be a Z-score of -2. So a common way to judge if a number is unusual (outlier) for normal data is to look at the Z-score.

Unusual High Values for Normal Data: $Z \geq +2$

Unusual Low Values for Normal Data: $Z \leq -2$



Hence, since the woman's Z-score was greater than or equal to +2, she is unusually tall compared to the women in the data set.

Example: The women's height data was normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. One woman in the data set was 57 inches tall and we said was unusually short. If you recall, her height was below the unusual low cutoff of 57.713 inches. What would be the Z-score for her height?

$$Z = \frac{(X \text{ value} - \text{Mean})}{\text{Standard Deviation}} = \frac{(57 - 63.195)}{2.741} \approx -2.26$$

Since the X-value is below the mean, the Z-score will be negative. So the woman that is 57 inches tall is 2.26 standard deviations below the mean. Remember if the Z-score is less than -2, it is unusually low. This confirms what we already knew.

Typical Z-scores: Remember that typical values are within one standard deviation from the mean. This would mean that typical Z-scores are between -1 and +1.

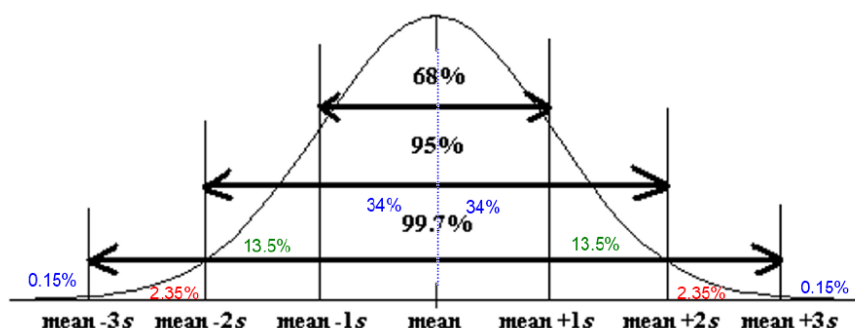
$$-1 \leq \text{Typical Z-scores} \leq +1$$

A woman with a height of 61 inches would have a Z-score of -0.80. Notice that this Z-score is between -1 and +1 on the number line. Therefore, 61 inches is a typical height for women in this data set.

Note: Not all values are typical or unusual. A person that is 1.5 standard deviations from the mean would be neither typical (Z-score not between -1 and +1) nor unusual (Z-score not greater than +2 or less than -2).

Empirical Rule

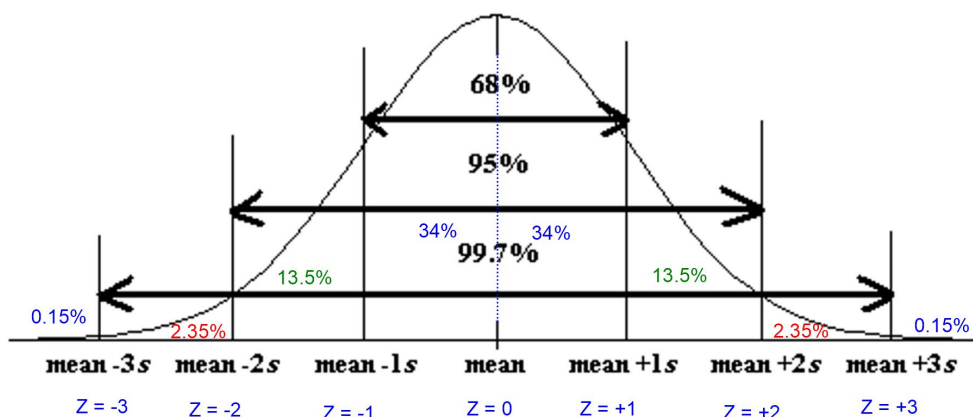
There is common percentages that go with normal (bell-shaped) data. Usually about 68% of normal data will be within one standard deviation of the mean (typical). About 95% of normal data will be within two standard deviations of the mean. About 99.7% of normal data will be within three standard deviations of the mean. These percentages are often referred to as the "Empirical Rule" or the "68-95-99.7 Rule".



Notice that we can use the 68%, 95% and 99.7% to figure out the sections. Since 68% makes up the middle two symmetric sections, we know each section is about 34%. Similarly, the middle four sections make up about 95%. Subtract out the middle two sections (68%) gives 27%. Divide that in half and you get two sections each making up 13.5% of the normal data. The middle six sections make up about 99.7%. Subtract out the middle four sections (95%) gives 4.7%. Divide that in half and you get two sections each making up 2.35% of the normal data. The end sections are calculated in a similar manner ($100\% - 99.7\% = 0.3\%$). Divide that into two symmetric tails and we get that each tail should be about 0.15%.

Remember the number of standard deviations from the mean is the Z-score. You can write the Z-scores for the bottom values in the Empirical rule. This is often called the "Standard Normal Curve". Notice the center of the curve is the mean (Z-score of zero) and the standard deviation of this curve is exactly one. When a computer program refers to a normal curve with a mean of zero and a standard deviation of one, they are talking about Z-scores and the Standard Normal Curve.





Many data sets are normal. We will see in the next chapter that many sampling distributions have a normal shape as well. It is therefore important to be able to calculate percentages associated with normal data and normal curves. Confidence Intervals and P-value are both extremely important topics that we will cover in chapter 3 and chapter 4 that involve the empirical rule and calculating percentages associated with normal curves.

Calculating Percentages for Normal Curves with StatKey

Computer software programs can calculate percentages associated with normal quantitative data. Go to www.lock5stat.com and click on "StatKey". Under the "Theoretical Distributions" menu click on "Normal". Notice the parameters are set at a mean of zero and a standard deviation of one. Remember this means it is set up to find Z-scores or to find percentages associated with Z-scores. The curve is sometimes called a "density curve". The idea is that the total area under the curve is 100%, so to find a percentage you find the area under the curve.

Normal Distribution

Mean	Standard Deviation
0	1

[Edit Parameters](#)

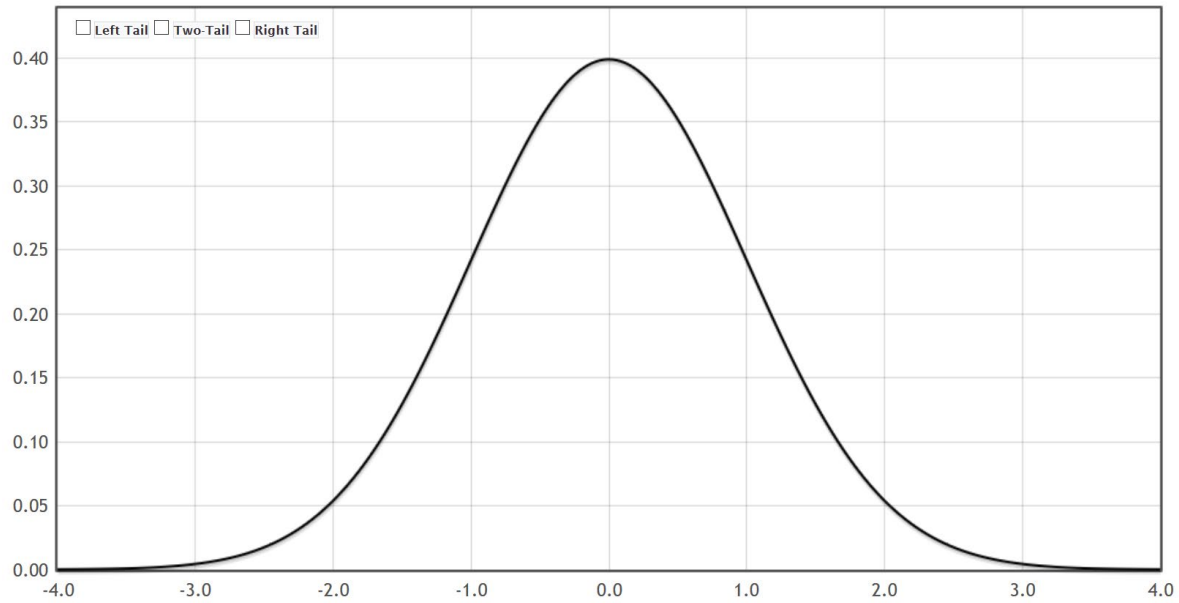


This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

StatKey Theoretical Distribution

Normal Distribution ▾

Reset Plot



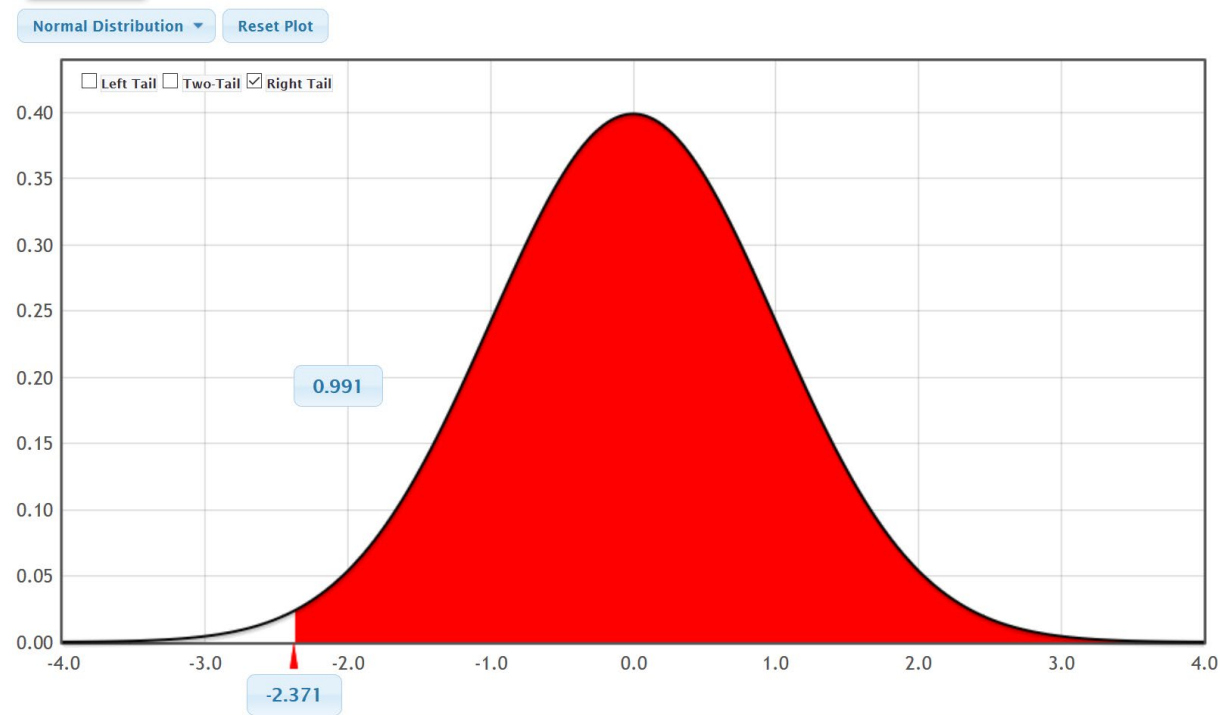
Notice that the curve has three buttons on the top left (Left Tail, Two-Tail, and Right Tail).



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Example: Suppose we want to find what percent of normal data has a Z-score of -2.371 or above. Since we are looking for above, click the right tail button. The upper box is the percentage and the lower box is the Z-score. In this case, we know the Z-score and are looking for the percentage. So in the bottom box type in -2.371 .

StatKey Theoretical Distribution

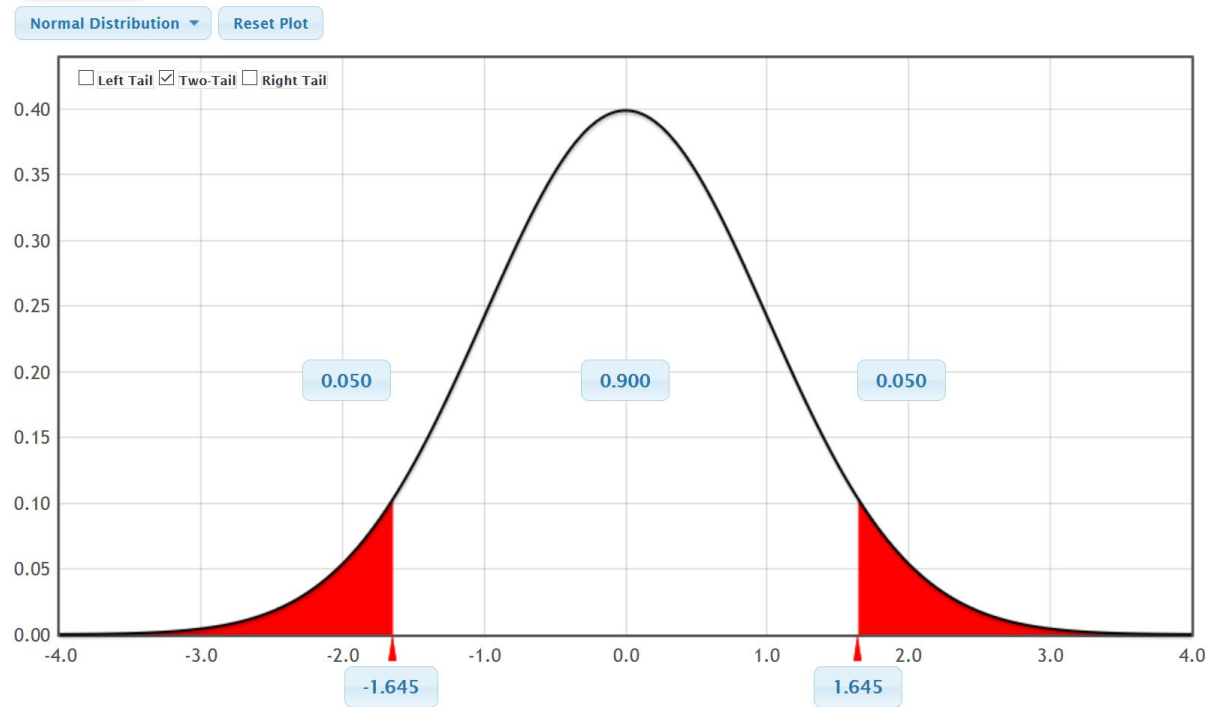


Notice the top box is the answer, 99.1% of normal data values will have a Z-score of -2.371 or higher.



Example: Push the “reset plot” button. Suppose we want to find the two Z-scores that 90% (0.9) of normal data values are in between. Since we are looking for “in between”, click the two-tail button. The upper boxes are the percentages and the lower boxes are the Z-scores. In this case, we know the percentage in between, but need to find the Z-scores. So in the upper middle box type in the decimal proportion equivalent of 90% (0.9).

StatKey Theoretical Distribution



Notice the Z-score answers we are looking for are at the bottom. Therefore, the middle 90% of normal data values have a Z-score between -1.645 and $+1.645$. These are the famous Z-scores for 90% confidence intervals that we will study in later chapters.

Percentages for any normal data

We often want to calculate percentages for normal quantitative data without calculating Z-scores first. StatKey can do that as well. Push the “reset plot” button. Right now, the mean is set at zero and the standard deviation is at one.

Normal Distribution

Mean	Standard Deviation
0	1

Edit Parameters



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Example: Suppose we want to calculate percentages associated with the women's height data we studied earlier. We found that the women's heights were normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. Click on the button that says, "edit parameters" and put those numbers into StatKey.

Normal Distribution

Mean	Standard Deviation
63.195	2.741

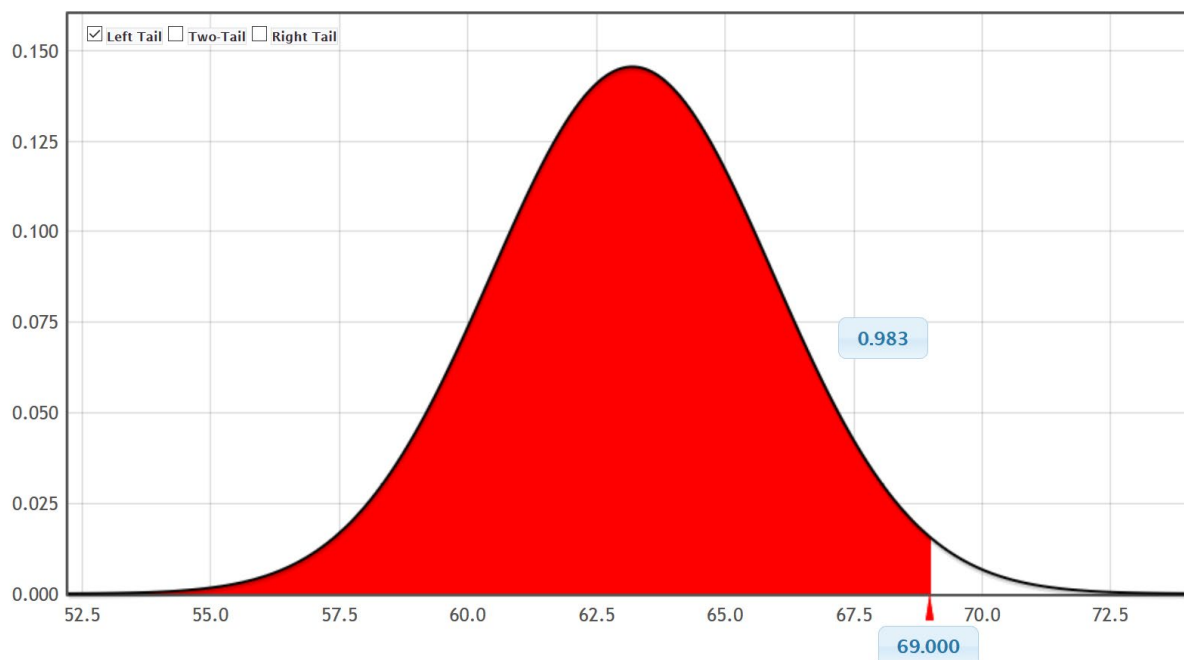
Edit Parameters

Suppose we want to know what percentage of women in the data have a height of 69 inches or less. Since we are looking for "less than", click on left tail. Remember the top box is the percentage (proportion). The bottom box is now the height. Since we know the height is 69, type in 69 into the bottom box. The proportion in the top box is our answer. So about 98.3% of the women in the sample data have a height below 69 inches.

StatKey Theoretical Distribution

Normal Distribution

Reset Plot

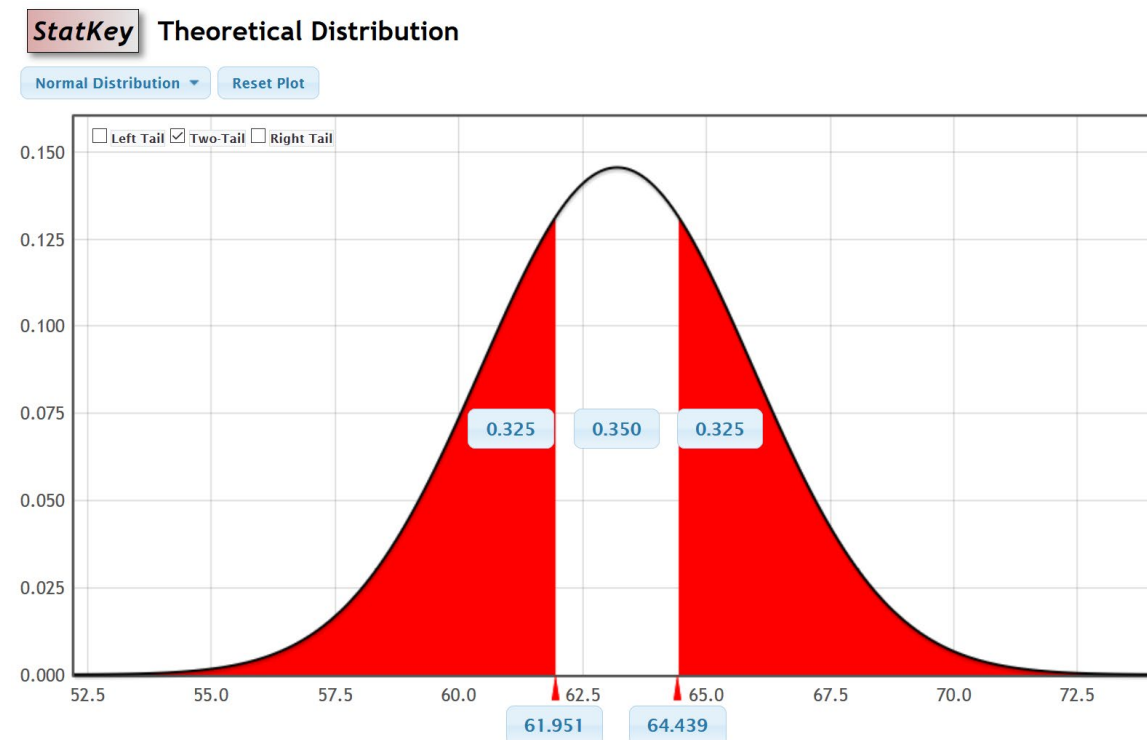


Note: Be careful about generalizing results of sample data to the population. This does not mean that 98.3% of all women have a height of 69 inches or below. As we learned in chapter one, samples may have bias and not represent the population.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Example: Suppose we wish to find the two heights corresponding to the middle 35%. That is the two heights that 35% of women are in between. Just push the “two-tail” button and put 0.35 in the upper middle box. The answer will be in the two lower boxes.



So about 35% of the women in the data have a height between 61.951 and 64.439 inches.

Note: These percentages are based on perfectly normal curves, yet real data is rarely perfectly normal. There are actually 15 women in the data had a height between 61.951 and 64.439. This was actually 37.5%. This is off from the theoretical percentage because the data was not perfectly normal. It is important to realize that theoretical distributions rarely match up exactly with real data.

Calculating Percentages for Normal Curves with Statcato

Z-scores, X-values and percentages for normal curves can also be calculated with Statcato. Go to the “Calculate” menu, click on “Probability Distributions” and then “Normal”.



Normal Probability Distribution

Help
F1

Distribution

Distribution Parameters:

Mean: 0

Standard deviation: 1

Compute:

☒ Probability density

☐ Cumulative probability

☐ Inverse cumulative probability

Inputs and Outputs

Input(s):

☐ Column:

☒ Constant:

Store Results in: (optional)

(e.g. C1 for column label, or variable name)

Compute
Close

If you leave the mean at zero and the standard deviation at one, then Statcato is set up to calculate Z-scores or percentages from Z-scores. To calculate a Z-score from a percentage less than the Z-score, put in the proportion (decimal equivalent of the percentage) into the box that says “constant”. Then click “inverse cumulative” and “compute”. For example, what is the Z-score that 85% of values in a normal data set are less than? The answer is under “X”. The Z-score is 1.0365.

Normal Distribution: mean = 0.0 stdev = 1.0

Input: 0.85


Type: Inverse cumulative probability

P(<=X) X

0.85 1.0365



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18


Normal Probability Distribution
×

Help
F1

Distribution

Distribution Parameters:

Mean:

Standard deviation:

Compute:

☐ Probability density
☐ Cumulative probability
☒ Inverse cumulative probability

Inputs and Outputs

Input(s):

☐ Column:

☒ Constant:

Store Results in: (optional)

(e.g. C1 for column label, or variable name)

Compute
Close

Suppose we want to find the percentage less than a Z-score of 2.36. Put 2.36 in the constant box and press "Cumulative Probability". The answer is under "P(<= X)". So the answer 0.990863 or about 99.1%.

Normal Distribution: mean = 0.0 stdev = 1.0

Input: 2.36

Type: Cumulative probability

X P(<=X)

2.36 0.990863



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Normal Probability Distribution

Help
F1

Distribution

Distribution Parameters:

Mean: 0

Standard deviation: 1

Compute:

☐ Probability density

☒ Cumulative probability

☐ Inverse cumulative probability

Inputs and Outputs

Input(s):

☐ Column:

☒ Constant: 2.36

Store Results in: (optional)

(e.g. C1 for column label, or variable name)

Compute
Close

We can also calculate X-values and percentage for those X-values for normally distributed data. We need to input the mean and standard deviation into Statcato. For example, earlier we saw some random sample data for women's heights was normally distributed with a mean of 63.195 and a standard deviation of 2.741. Suppose we want to find the percentage of women in the data that have a height below 64 inches. We see that the answer is 0.615502 or about 61.6%. Note that Statcato can only calculate for less than. If we want to know what percent of women in the data have a height above 64 inches, we first calculate less than and then subtract the answer from 100%. In this case, $100\% - 61.6\% = 38.4\%$.

Normal Distribution: mean = 63.195 stdev = 2.741


Input: 64.0

Type: Cumulative probability

X $P(\leq X)$

64.0 0.615502



 Normal Probability Distribution
 ×

Help
 F1

Distribution

Distribution Parameters:

Mean:

Standard deviation:

Compute:

☐ Probability density

☒ Cumulative probability

☐ Inverse cumulative probability

Inputs and Outputs

Input(s):

☐ Column:

☒ Constant:

Store Results in: (optional)

(e.g. C1 for column label, or variable name)

You can also use the “Inverse Cumulative Probability” function to calculate the height that 15% of women are taller than. Remember, Statcato only works with less than, so if 15% of women are greater than this height, than 85% of women are less than this same height. So we will enter 85% (0.85) into the constant box. We see the answer under “X”. Therefore, 85% of women have a height less than 66.0358 inches. This also means that 15% of women have a height above 66.0358 inches.

Normal Distribution: mean = 63.195 stdev = 2.741

Input: 0.85

Type: Inverse cumulative probability

$P(\leq X)$ X

0.85 66.0358



This chapter is from *Introduction to Statistics for Community College Students*,
 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
 under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Normal Probability Distribution ×

Help F1

Distribution

Distribution Parameters:

Mean:

Standard deviation:

Compute:

☐ Probability density

☐ Cumulative probability

☒ Inverse cumulative probability

Inputs and Outputs

Input(s):

☐ Column:

☒ Constant:

Store Results in: (optional)

(e.g. C1 for column label, or variable name)

Calculating between is challenging with Statcato. Statcato does not have a between button, so we must use the percentages less than an X-value. If we want to find the two values that the middle 40% are in between, we have to think about the percentages less than each X-value. If 40% is in the middle, then the remaining 60% is divided into the two tails. Therefore, each tail must be 30%. So the X-value on the left will have 30% (0.3) less than. The X-value on the right will have 70% (0.7) less than. Put 0.3 into the "Constant" box and press inverse cumulative. Then put 0.7 into the "Constant" box and press inverse cumulative. For women's heights, we would get that the middle 40% of women's heights are between 61.7576 inches and 64.6324 inches.

Normal Distribution: mean = 63.195 stdev = 2.741

Input: 0.3

Type: Inverse cumulative probability

$P(\leq X)$ X

0.3 61.7576

Normal Distribution: mean = 63.195 stdev = 2.741

Input: 0.7

Type: Inverse cumulative probability

$P(\leq X)$ X

0.7 64.6324



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Note on Rounding Statistics for Quantitative Data

It is often best to not round if you are unsure. Data analysts usually prefer better accuracy and can round to their own specifications. Rounding too much interferes with accuracy. If you must round, here are some general guidelines.

Percentages and proportions are usually rounded to three significant figures. Proportions are rounded to the thousandths place and percentages are rounded to the tenths place.

Quantitative statistics like the mean or standard deviation are usually rounded to one more decimal place to the right than the original data has. Notice the women's heights data is rounded to the tenths place (one number to the right of the decimal). So statistics calculated from this data would usually be rounded to the hundredths place (two numbers to the right of the decimal).

Mean (women's height) = $63.195 \approx 63.20$ inches

Standard Deviation (women's height) = $2.741 \approx 2.74$ inches

B
Women Ht (in)
64.3
66.4
62.3
62.3
59.6
63.6

Practice Problems Section 1F

1. Answer the following questions:

- What is meant by saying that data is normally distributed or "normal"?
- Define the mean average and explain how it is calculated.
- Define the standard deviation and explain how it is calculated.

2. Answer the following questions:

- If a data set is normally distributed, what measure of average should we use?
- If a data set is normally distributed, what measure of spread should we use?
- If a data set is normally distributed, how many standard deviations from the mean is considered typical?
- If a data set is normally distributed, what is the formula for finding typical values?
- If a data set is normally distributed, approximately what percentage is typical?
- If a data set is normally distributed, how many standard deviations from the mean is considered unusual?
- If a data set is normally distributed, approximately what percentage of the data is unusually high?
- If a data set is normally distributed, approximately what percentage of the data is unusually low?



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

(#3-4) Directions: The following graphs and statistics were calculated with Statcato and the “Bear” data from the website www.matt-teachout.org. Use the dot plot, histogram and summary statistics to answer the following questions. Here are the formulas for typical and unusual values.

Mean – Standard Deviation ≤ Typical Values for Normal Data ≤ Mean + Standard Deviation

Unusual Low Cutoff for Normal Data = Mean – (2 × Standard Deviation)

Unusual High Cutoff for Normal Data = Mean + (2 × Standard Deviation)

3. Bear neck circumference (inches)

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- Is the data set normally distributed? (Yes or No)
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? (Give the number and the name of the statistic used.)
- How much typical spread does the data set have? (Give the number and the name of the statistic used.)
- Find two numbers that typical values fall in between.
- What is the unusual high (high outlier) cutoff for this data?
- What is the unusual low (low outlier) cutoff for this data?
- List all high outliers in this data set. If there are no high outliers, put “none”.
- List all low outliers in this data set. If there are no high outliers, put “none”.

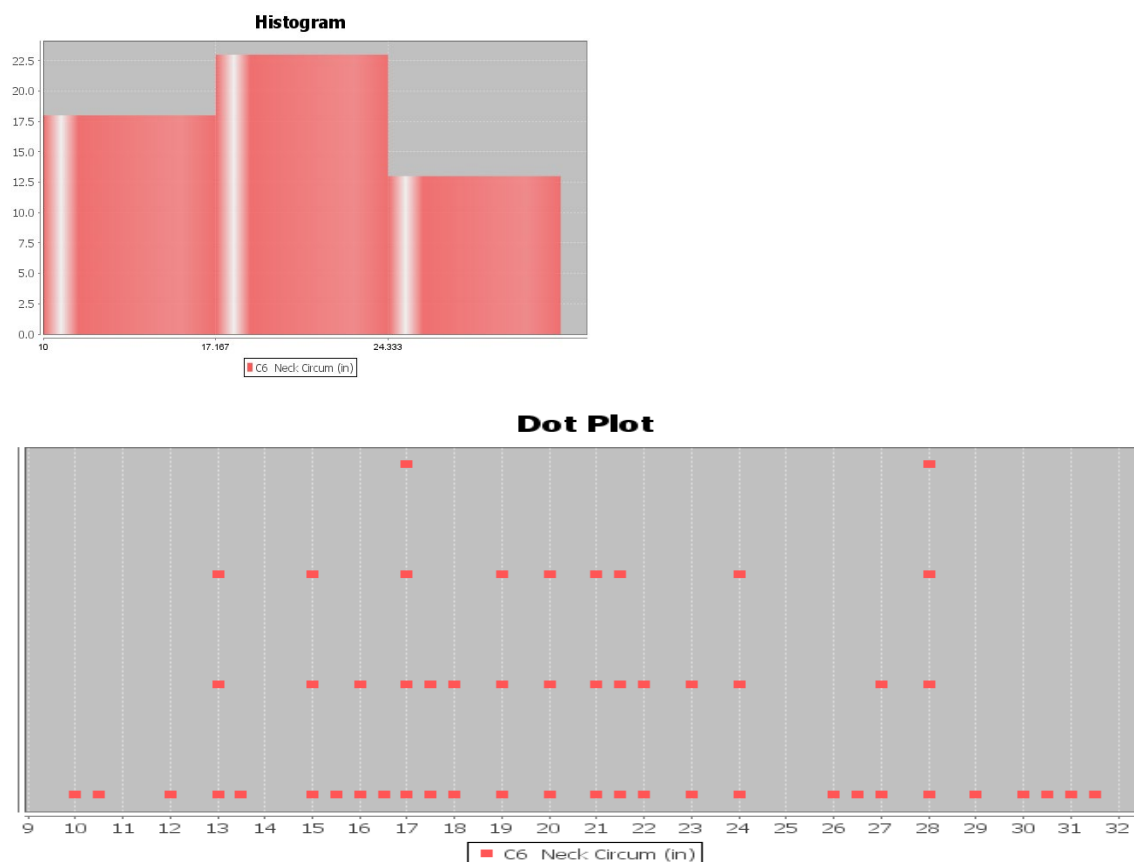
Descriptive Statistics

Variable	Mean	Standard Deviation
C6 Neck Circum (in)	20.556	5.641

Variable	Min	Max
C6 Neck Circum (in)	10.0	31.5

Variable	N total
C6 Neck Circum (in)	54





4. Bear Chest Size (inches)

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- Is the data set normally distributed? (Yes or No)
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? (*Give the number and the name of the statistic used.*)
- How much typical spread does the data set have? (*Give the number and the name of the statistic used.*)
- Find two numbers that typical values fall in between.
- What is the unusual high (high outlier) cutoff for this data?
- What is the unusual low (low outlier) cutoff for this data?
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".

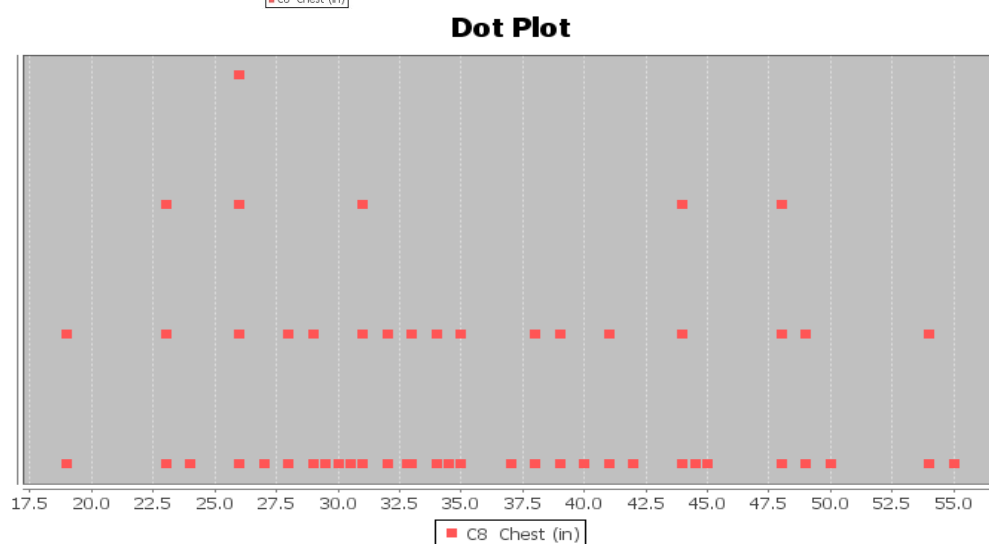
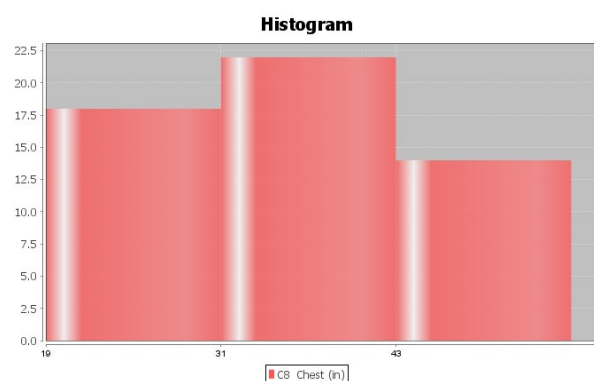


Descriptive Statistics

Variable	Mean	Standard Deviation
C8 Chest (in)	35.663	9.352

Variable	Min	Max
C8 Chest (in)	19.0	55.0

Variable	N total
C8 Chest (in)	54



(#5-8) Directions: Open the “Health” data from the website www.matt-teachout.org. (Look under “Statistics” tab and then click the “data sets” tab.) Go to www.lock5stat.com and open StatKey. Under the “Descriptive Statistics and Graphs” menu, click on “One Quantitative Variable”. Paste the indicated column of quantitative data into StatKey, create a dot plot and histogram, and find the summary statistics. Then answer the following questions. Here are the formulas for typical and unusual values.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Mean – Standard Deviation \leq Typical Values for Normal Data \leq Mean + Standard Deviation

Unusual Low Cutoff for Normal Data = Mean – (2 \times Standard Deviation)

Unusual High Cutoff for Normal Data = Mean + (2 \times Standard Deviation)

5. Women's Diastolic Blood Pressure (Millimeters of Mercury (mm of Hg))

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- Is the data set normally distributed? (Yes or No)
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? *(Give the number and the name of the statistic used.)*
- How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- Find two numbers that typical values fall in between.
- What is the unusual high (high outlier) cutoff for this data?
- What is the unusual low (low outlier) cutoff for this data?
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".

6. Women's Wrist Circumference (Inches)

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- Is the data set normally distributed? (Yes or No)
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? *(Give the number and the name of the statistic used.)*
- How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- Find two numbers that typical values fall in between.
- What is the unusual high (high outlier) cutoff for this data?
- What is the unusual low (low outlier) cutoff for this data?
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".

7. Men's Height (Inches)

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- Is the data set normally distributed? (Yes or No)
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? *(Give the number and the name of the statistic used.)*
- How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- Find two numbers that typical values fall in between.
- What is the unusual high (high outlier) cutoff for this data?
- What is the unusual low (low outlier) cutoff for this data?
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".



8. Men's Weight (Pounds)

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- Is the data set normally distributed? (Yes or No)
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? *(Give the number and the name of the statistic used.)*
- How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- Find two numbers that typical values fall in between.
- What is the unusual high (high outlier) cutoff for this data?
- What is the unusual low (low outlier) cutoff for this data?
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".

(#9-18) *Directions: Use the following formula when needed and answer the following questions about Z-scores.*

$$Z = \frac{(\text{Amount} - \text{Mean})}{\text{Standard Deviation}}$$

9. Write the definition of a Z-score.

10. Explain how we can use Z-scores to tell if a number is typical in normal data?

11. Explain how we can use Z-scores to tell if a number is unusual in normal data?

12. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Bud has an IQ of 143. Use this information to answer the following Z-score questions.

- Calculate the Z-score for Bud's IQ.
- Write a sentence to explain the Z-score in context.
- Is Buds' IQ unusually high compared to other people in the data set? Explain your answer.

13. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Jan has an IQ of 89. Use this information to answer the following Z-score questions.

- Calculate the Z-score for Jan's' IQ.
- Write a sentence to explain the Z-score in context.
- Is Jan's' IQ unusually low compared to other people in the data set? Explain your answer.

14. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of \$46.89 and a standard deviation of \$12.44. Maria spent \$105.12 on merchandise in the store. Use this information to answer the following Z-score questions.

- Calculate the Z-score for the amount Maria spent.
- Write a sentence to explain the Z-score in context.
- Is the amount Maria spent unusually high compared to other people in the data set?
Explain your answer.



15. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of \$46.89 and a standard deviation of \$12.44. Julie spent \$13.61 on merchandise in the store. Use this information to answer the following Z-score questions.

- a) Calculate the Z-score for the amount Julie spent.
- b) Write a sentence to explain the Z-score in context.
- c) Is the amount Julie spent unusually low compared to other people in the data set?
Explain your answer.

16. Neck circumferences of bears are normally distributed with a mean circumference of 20.556 inches and a standard deviation of 5.641 inches. A bear has a neck circumference of 13.7 inches.

- a) Calculate the Z-score for this bears neck circumference.
- b) Write a sentence to explain the Z-score in context.
- c) Is this bears' neck circumference unusually low compared to other bears in the data set?
Explain your answer.

17. Chest sizes of bears was normally distributed with a mean chest size of 35.663 inches and a standard deviation of 9.352 inches. A bear has a chest size of 57 inches.

- a) Calculate the Z-score for this bears chest size.
- b) Write a sentence to explain the Z-score in context.
- c) Is this bears' chest size unusually large compared to other bears in the data set?
Explain your answer.

18. The diastolic blood pressure of a random sample of women had a mean of 67.425 mm of Hg and a standard deviation of 11.626. A woman in the data has a diastolic blood pressure of 72 mm of Hg.

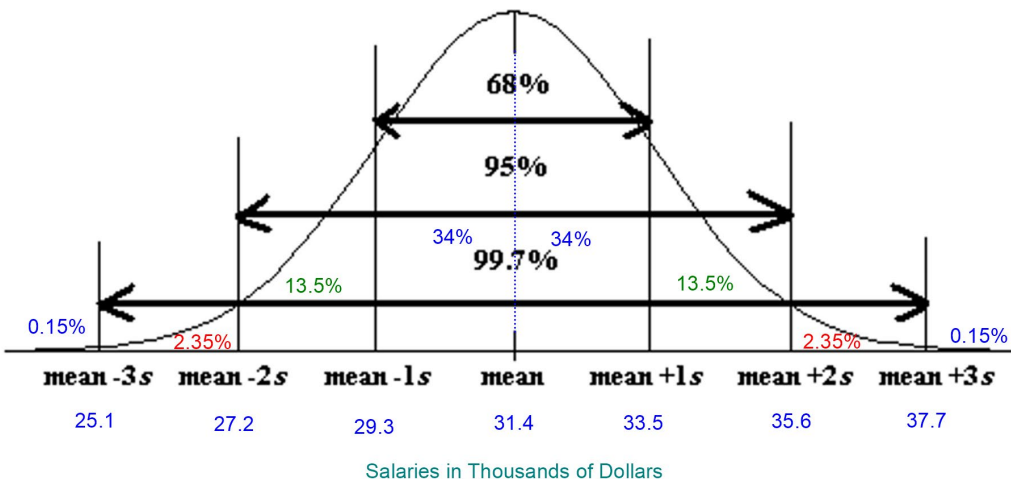
- a) Calculate the Z-score for this woman's diastolic blood pressure.
- b) Write a sentence to explain the Z-score in context.
- c) Is this woman's diastolic blood pressure unusually high compared to other women in the data set?
Explain your answer.

(#19-25) Answer the following questions about the empirical rule.

19. Draw that standard normal curve. Label the mean and the values for one, two and three standard deviations above and below the mean. Also, label the percentages that make up the empirical rule.



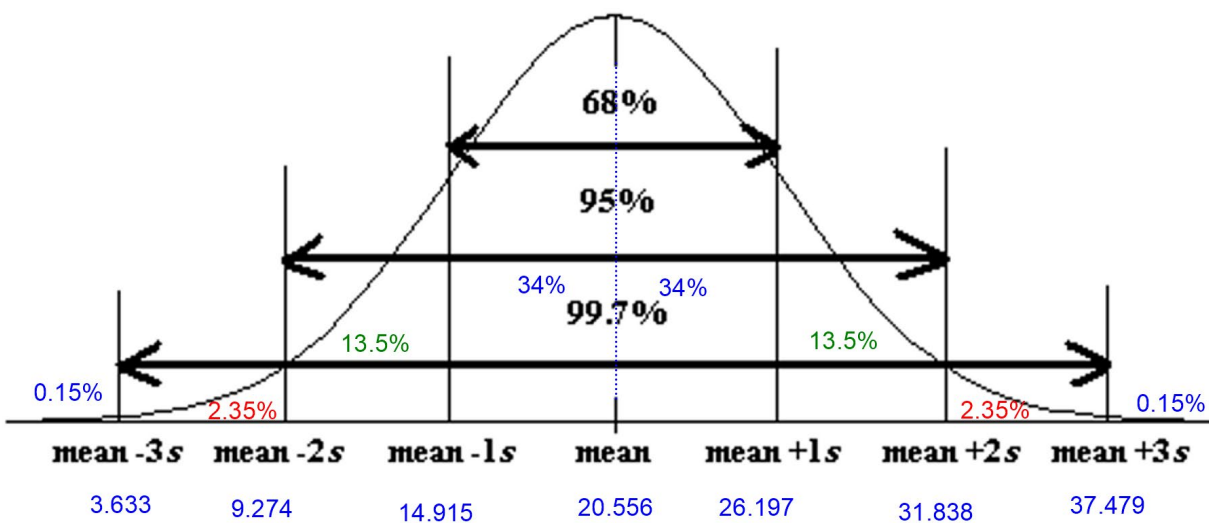
20. The salaries of employees at a company are normally distributed with a mean of 31.4 thousand dollars and a standard deviation of 2.1 thousand dollars. Use the Empirical Rule graph below to answer the following questions.



- What percentage of the employees have a salary between 27.2 thousand dollars and 35.6 thousand dollars?
- What percentage of the employees have a salary between 29.3 thousand dollars and 33.5 thousand dollars?
- What percentage of the employees have a salary between 25.1 thousand dollars and 37.7 thousand dollars?
- What percentage of the employees have a salary greater than 33.5 thousand dollars?
- What percentage of the employees have a salary less than 27.2 thousand dollars?
- Typical values for a normal curve are one standard deviation from the mean. Find two salaries that typical employee salaries fall in between?
- The unusual high cutoff is two standard deviations above the mean. What salary represents the unusual high cutoff, which is the salary that 2.5% of the employees are greater than?
- The unusual low cutoff is two standard deviations below the mean. What salary represents the unusual low cutoff, that is the salary that 2.5% of the employees are less than?



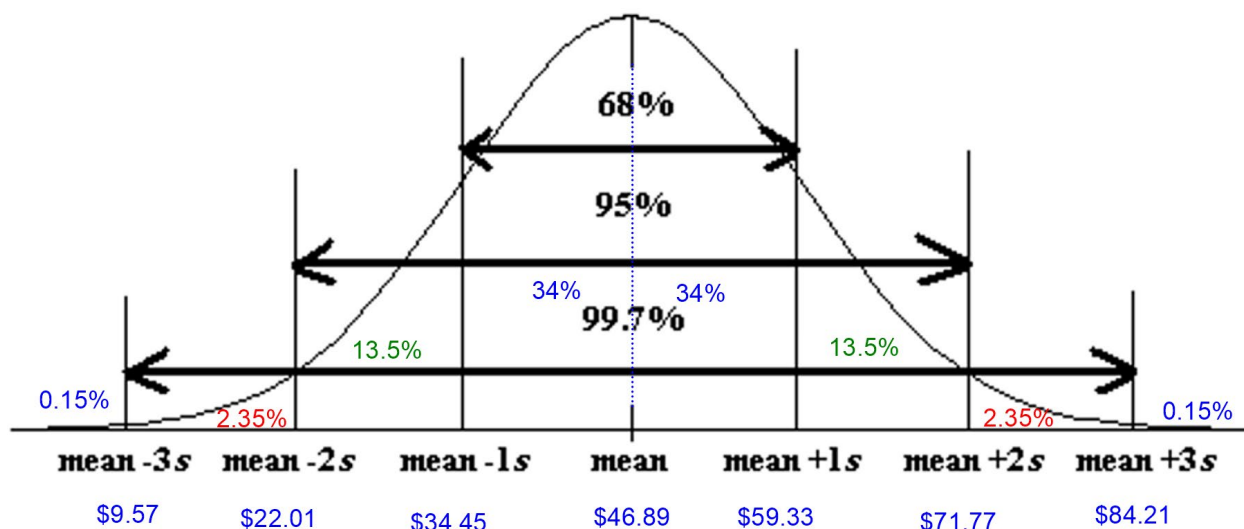
21. Neck circumferences of a sample of bears are normally distributed with a mean circumference of 20.556 inches and a standard deviation of 5.641 inches. Use the Empirical Rule graph below to answer the following questions.



- What percent of the bears have a neck circumference between 14.915 inches and 31.838 inches?
- What percent of the bears have a neck circumference less than 26.197 inches?
- Typical bears have a neck circumference between what two amounts?
- What is the unusual high cutoff, that is the bear neck circumference that 2.5% of bears are more than.
- What is the unusual low cutoff, that is the bear neck circumference that 2.5% of bears are less than.
- What is the bear neck circumference that 84% of bear neck circumferences are more than?
- What percent of the bears have a neck circumference less than 14.915 inches?



22. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of \$46.89 and a standard deviation of \$12.44. Use the Empirical Rule graph below to answer the following questions.



- What percent of customers spent between \$71.77 and \$84.21 in the store?
- What percent of customers spent less than \$34.45 in the store?
- Typical customers spent between what two amounts?
- What is the unusual high cutoff, that is the dollar amount that 2.5% of customers spent more than.
- What is the unusual low cutoff, that is the dollar amount that 2.5% of customers spent less than.
- What is the dollar amount that 16% of customers spent more than?
- What percent of customers spent less than \$71.77?

23. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Use this information to answer the following questions. Go to www.lock5stat.com and open StatKey. Under the "Theoretical Distributions" menu, click on "Normal".

- Use StatKey to calculate what percent of people in the IQ sample data that have an IQ greater than 77.
- Use StatKey to calculate what percent of people in the IQ sample data that have an IQ less than 108.
- Use StatKey to calculate what percent of people in the IQ sample data that have an IQ between 95 and 120.
- Use StatKey to find the IQ score that 60% of people are less than.
- Use StatKey to find the IQ score that 85% of people are greater than.
- Use StatKey to find two IQ scores that the middle 40% of people are in between.



24. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of \$46.89 and a standard deviation of \$12.44. Go to www.lock5stat.com and open StatKey. Under the “Theoretical Distributions” menu, click on “Normal”.

- Use StatKey to calculate the percent of people that spent more than \$25.
- Use StatKey to calculate the percent of people that spent less than \$50.
- Use StatKey to calculate the percent of people spent between \$35 and \$60.
- Use StatKey to find the amount of money spent that 37% of people are less than.
- Use StatKey to find the amount of money spent that 15% of people are more than.
- Use StatKey to find two amounts that the middle 60% of people are in between.

25. The diastolic blood pressure of a random sample of women had a mean of 67.425 mm of Hg and a standard deviation of 11.626. Go to www.lock5stat.com and open StatKey. Under the “Theoretical Distributions” menu, click on “Normal”.

- Use StatKey to calculate the percent of women that have a diastolic blood pressure below 75 mm of Hg.
- Use StatKey to calculate the percent of women that have a diastolic blood pressure above 50 mm of Hg.
- Use StatKey to calculate the percent of women that have a diastolic blood pressure between 60 and 70 mm of Hg.
- Use StatKey to find the diastolic blood pressure that 80% of women are lower than.
- Use StatKey to find the diastolic blood pressure that 45% of women are higher than.
- Use StatKey to find the two diastolic blood pressures that the middle 75% of women are in between.

Section 1G – Quantitative Data Analysis for Non-Normal Data and Summary Statistics

Vocabulary

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Normal Data: Data that is bell shaped, symmetric and unimodal.

Skewed Right Data: Also called positively skewed. Data where the center is on the far left and has a long tail to the right.

Skewed Left Data: Also called negatively skewed. Data where the center is on the far right and has a long tail to the left.

Sample Size: Also called the total frequency. The number of values are in a data set.

Median Average: The center of the data when the numbers are put in order. Also called the “50th Percentile” (P_{50}). Since about 50% of the numbers in the data set are less than the median. It is also called the “Second Quartile” (Q_2). The average for a data set that is not normal.

First Quartile (Q_1): The number that about 25% of the data values are less than. Used for typical values for data that is not normal.

Third Quartile (Q_3): The number that about 75% of the data values are less than. Used for typical values for data that is not normal.

Interquartile Range (IQR): The distance between the middle 50% of the numbers in a data set. Calculated by subtracting the 1st and 3rd quartiles. The measure of typical spread for a data set that is not normal.

Maximum: The largest number in a data set.



*This chapter is from **Introduction to Statistics for Community College Students**, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Minimum: The smallest number in a data set.

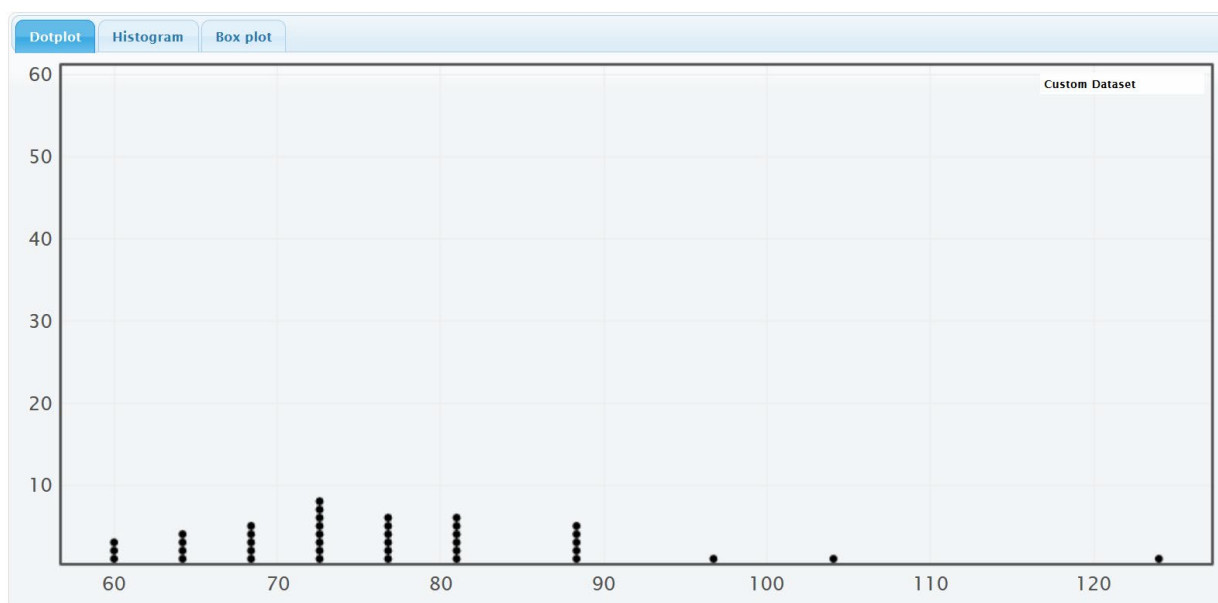
Range: A quick measure of total spread. Calculated by subtracting the minimum and maximum values in a data set.

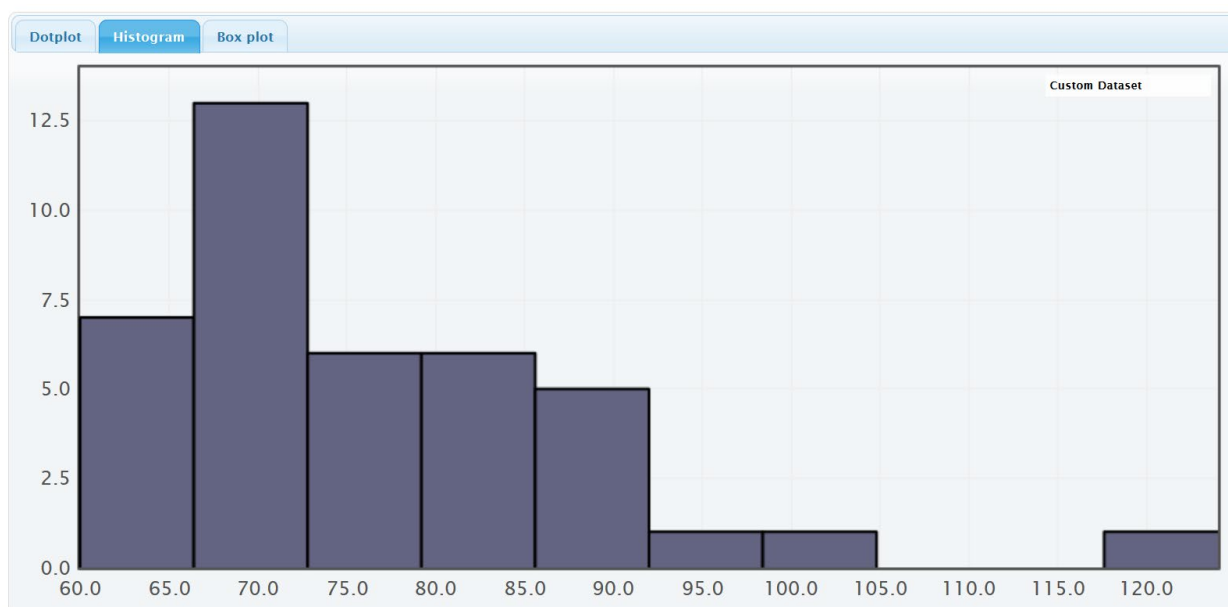
Outliers: Unusual values in the data set.

Introduction

When a data set is normal (or bell-shaped), we use the mean as our average and the standard deviation as our measure of typical spread. Not all data sets are normal though. Let us explore some data that is not normally distributed.

Let us look at another example from the health data. This time we will look at women's pulse rates in beats per minute (BPM). Go to www.matt-teachout.org and click on the "Statistics" tab and then the "Data Sets" tab. Open the health data in Excel. Copy the women's pulse rate data. Now go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable". Under "Edit Data", paste the women's pulse rate data into StatKey. Uncheck the box that says, "First column is identifier". Check the box that says, "Data has header row". Push "OK". Here are the graphs and summary statistics.





Notice first that this is not normal data. The highest bar (center) is on the far left. The graph has a short tail to the left of the highest bar and a long tail to the right of the highest bar. This shape is called “skewed right” or “positively skewed”. We can adjust the number of bars (buckets) by using the slider on the right of the graph.

Summary Statistics

Statistic	Value
Sample Size	40
Mean	76.300
Standard Deviation	12.499
Minimum	60
Q_1	68.000
Median	74.000
Q_3	80.000
Maximum	124

Remember the mean and standard deviation are only accurate if the data is normal. Therefore, for this data set, we should not use the mean as the average and we should not use the standard deviation as our typical spread.

So what statistics should we use? Here is the general rule for skewed data or any data that is not normal.

Summary statistics for non-normal data

Average: Median

Typical Spread: Interquartile Range (IQR)

Typical Values: Between the first quartile (Q_1) and the third quartile (Q_3)

Outliers: Boxplot will indicate if there are outliers.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Quartiles are based on the numbers in order, so are much more accurate for data that is not normally distributed. The median is also called the 2nd quartile or the 50th percentile. It is the center of the data when the numbers are in order. About 50% of the numbers will be less than the median and about 50% of the numbers will be greater than the median. When a data set is not normally distributed, we use the median as our average. It is much closer to the center. Look at the histogram above. The summary statistics provided by StatKey show us that the mean was 76.3 beats per minute (bpm) and the median was 74 bpm. Notice 74 is closer to the highest bar in the data set. In other words, the median is closer to the center and a more accurate average than the mean. Mean averages are based on distances so will be pulled off the center in the direction of the skew.

The median is calculated by first putting the numbers in order from smallest to largest. If there is one number in the middle (sample size n is odd), then that is the median. If there are two numbers in the middle (sample size n is even), then the median will be half way between the two numbers in the middle.

The first quartile (Q_1) is also called the 25th percentile and is the number that about 25% of the data is less than. The third quartile (Q_3) is also called the 75th percentile and is the number that about 75% of the data is less than. The first and third quartiles are markers that mark the middle 50% of the data when it is in order. The middle 50% is considered "typical" in a data set that is not normally distributed. For normal data, we want the middle 68% (empirical rule) because there is more data in the middle.

The distance between the first and third quartiles is called the interquartile range (IQR). This is the best measure of typical spread for data that is not normally distributed. StatKey does not list the IQR in its summary statistics, but we can calculate it with the following formula.

$$\text{IQR} = Q_3 - Q_1$$

Since our women's pulse rate data was skewed right, we would use the following statistics.

Variable and Units: Women's pulse rates in beats per minute (bpm)

Minimum: The lowest pulse rate for these women was 60 bpm.

Maximum: The highest pulse rate for these women was 124 bpm.

Average: The average pulse rate for these women is 74 bpm (median).

Typical spread: $\text{IQR} = Q_3 - Q_1 = 80 - 68 = 12$ bpm

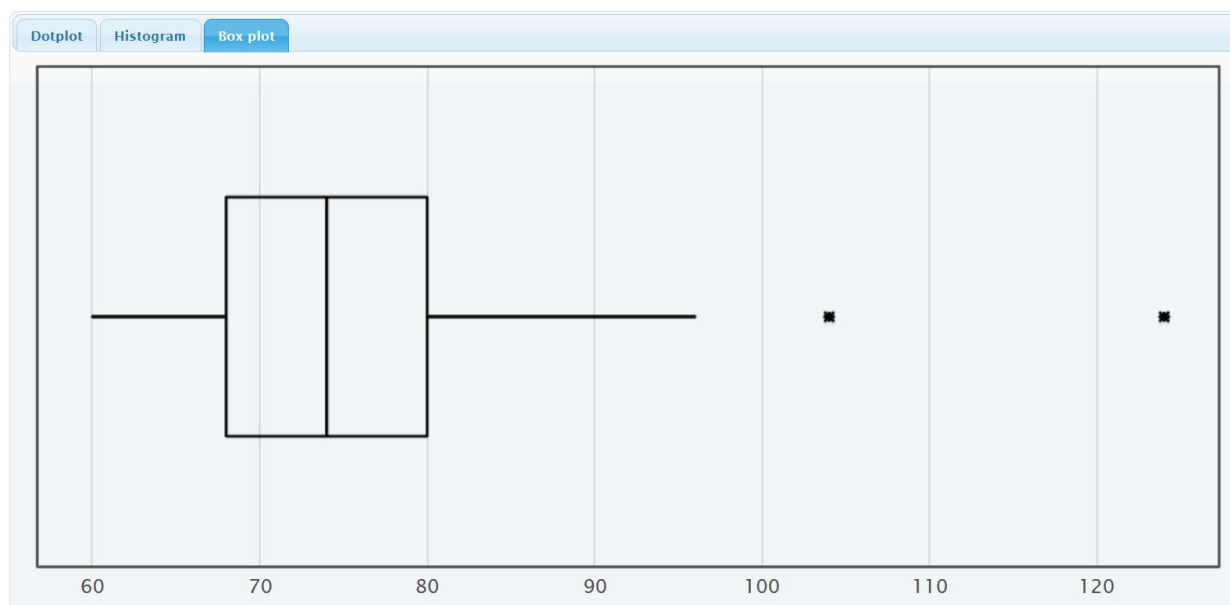
Typical women in the data set had a pulse rate within 12 bpm of each other.

Typical Values: Typical pulse rates are between 68 bpm (Q_1) and 80 bpm (Q_3).



Finding outliers for non-normal data

To find outliers for data sets that are not normally distributed, we will introduce another graph. The graph is called a “box and whisker plot” or “box plot” for short.



A box plot is a graph of the first quartile, median, third quartile and outliers. It is the perfect graph to look at when a data set is not normal. The left of the box is Q_1 (68 bpm) and far right of the box is Q_3 (80 bpm). So the box represents the typical values (middle 50%). The line inside the box is the median average of 74 bpm. The lines that go to the left and right of the box are called whiskers. The whiskers go to the lowest and highest numbers in the data set that are not unusual (not outliers). The outliers are usually denoted by stars in StatKey and circles and triangles in Statcato. See the two stars the far right. Those are both outliers. There are two unusually high pulse rates in the data set. In StatKey, you can hold your cursor over the stars and they will tell you what the numbers are. In this case, the two high outliers are at 104 bpm and 124 bpm. There are no unusually low values since we do not see any stars on the left of the graph.

In case you are wondering, here are the formulas used by computer programs to determine outliers in a box plot. You do not need to calculate this yourself. The computer has already found your unusual values.

Unusual high (high outlier) cutoff: $Q_3 + (1.5IQR)$

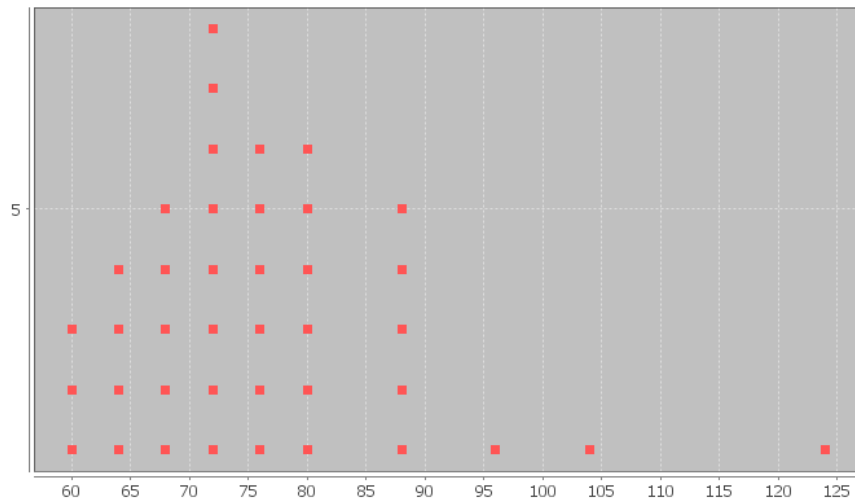
Unusual low (low outlier) cutoff: $Q_1 - (1.5IQR)$

Note about box plots and normal data: Remember, a box plot is a graph of the quartiles and the median. They work really well for data that is not normal. However, they do not show the mean or standard deviation, so it is important to be careful how you interpret box plots for normal data. Normal data has different characteristics than those shown on a box plot. For example, typical values for normal data are not between Q_1 and Q_3 . In addition, the outlier cutoffs are different for normal data so there may be differences in what is considered an outlier.

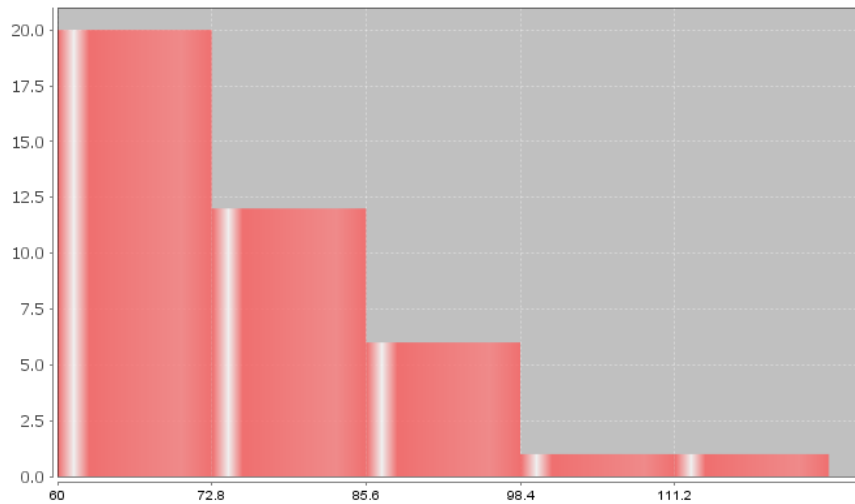
In the last section, we saw that we could also calculate dot plots, histograms, box plots and summary statistics with Statcato. Copy and paste the data into a column of Statcato. Then go to the graph menu and click on “dot plot”, “histogram” or “box plot”.



Dot Plot of Women's Pulse Rates (Beats Per Minute)



Histogram of Women's Pulse Rates (Beats Per Minute)





Notice that something is wrong with the Statcato box plot. The outliers have been left off. This is a common problem. To fix this, right click on the box-plot. Click on “zoom out” and “range axis”. You may have to do this multiple times. You want to be able to see the minimum value (60 bpm) and maximum value (124 bpm) on the scale of the graph. Here is the correct box plot.

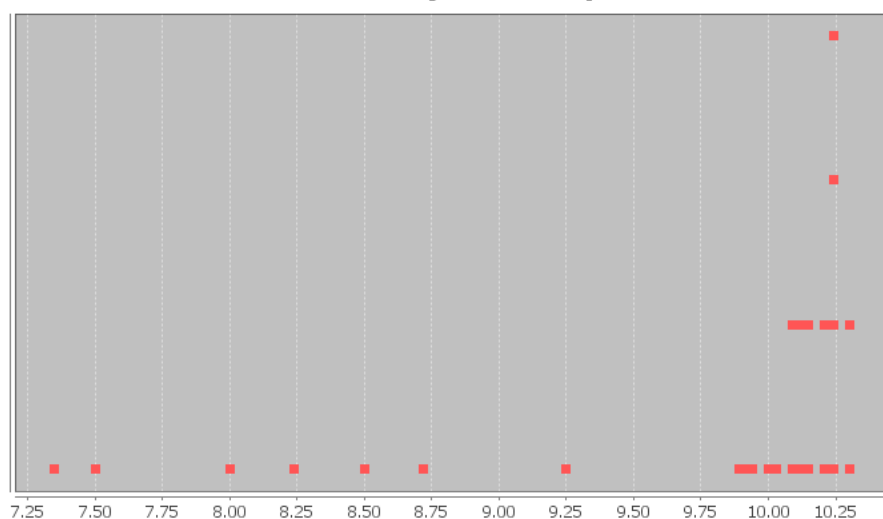


Notice Statcato designated 104 with a circle (regular outlier) and 124 with a triangle (far out outlier). The dot in the middle of the box plot is the mean. Most box plots do not have the mean, but Statcato puts it in so that you can compare it to the median.

Let us look at some other examples.

Here is some salary data from a small company with 26 employees. The salaries are given in dollars per hour. We created a dot plot and histogram for this data.

Dot Plot of Sallary in Dollars per Hour

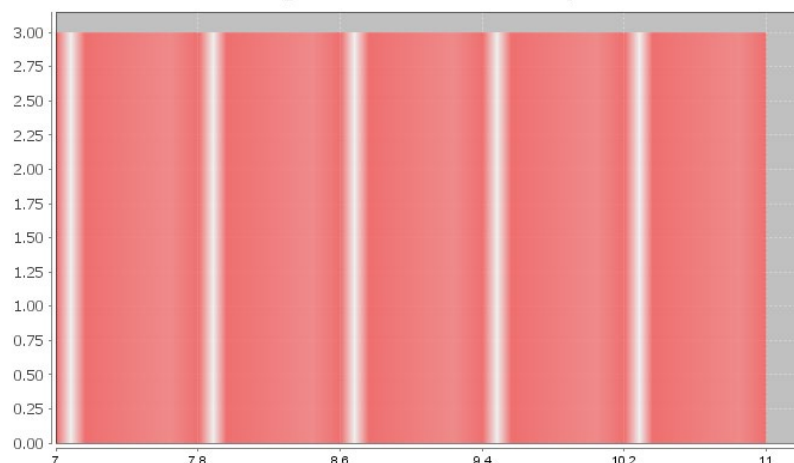


the far right and I have two bars to the right of the highest hill and seven bars to the left of the highest hill, I would classify that as skewed left. Some call this “negatively skewed” since negative numbers are to the left on the number line.

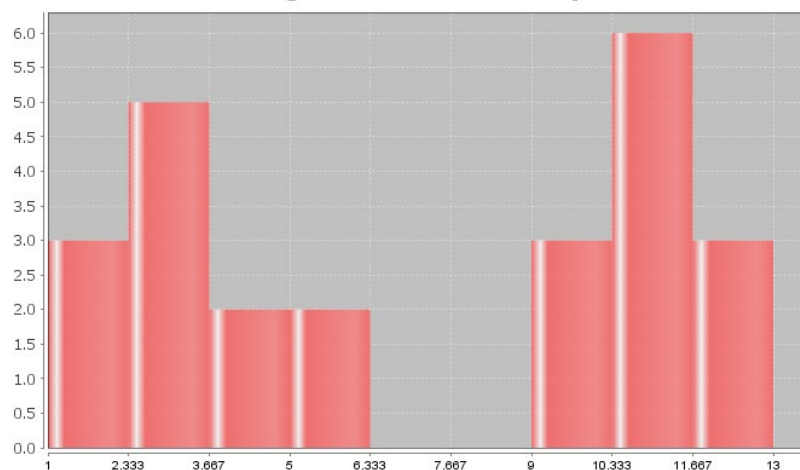
Here are a couple unusual shapes that sometimes appear.

A graph that looks like a rectangle is called “uniform”. A graph with two distinct high bars is called “bimodal”.

Histogram with Uniform Shape



Histogram with Bimodal Shape



Summary Statistics: Measures of Center, Spread and Position

Though the mean, median, standard deviation and IQR are used most often in data analysis, there are many different types of statistics that can be used to dig deeper into the data. We will not be covering these statistics in depth, but it is good to at least have an idea of what they measure.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Measures of Center

Mean Average: The balancing point in terms of distances. The measure of center or average used when a data set is bell shaped (normal).

Median Average: The center of the data in terms of order. Also called the second quartile (Q2) or the 50th percentile. Approximately 50% of the data will be less than the median and 50% will be above the median. This is the measure of center or average used when a data set is skewed (not bell shaped).

Mode: The number that occurs most often in a data set. Data sets may have no mode, one mode, or multiple modes. It is also sometimes used in bimodal or multimodal data.

Midrange: A quick measure of center that is usually not very accurate, but can be calculated quickly without a computer. $(\text{Max} + \text{Min}) / 2$

Measures of Spread

Standard Deviation: How far typical values are from the mean in a bell shaped data set. It is the most accurate measure of spread for bell shaped data. If you add and subtract the mean and standard deviation, you get two numbers that typical values in a bell shaped data set fall in between. It can also be used to find unusual values in bell shaped data. Should not be used unless the data is bell shaped.

Variance: The standard deviation squared. A measure of spread used in ANOVA testing. Only accurate when the data is bell shaped.

Range: A quick measure of spread that is not very accurate. It is based on unusual values and does not measure typical values in the data set. It can be calculated quickly without a computer. $(\text{Max} - \text{Min})$

Interquartile range (IQR): How far typical values are from each other in a skewed data set. Measures the length of the middle 50% of the data. It is the most accurate measure of spread for skewed data sets. Should not be used when data is bell shaped. $(Q3 - Q1)$

Measures of Position

Minimum: The smallest number in the data set. Is sometimes classified as an unusual value (outlier).

Maximum: The largest number in the data set. Is sometimes classified as an unusual value (outlier).

First Quartile (Q1): The number that approximately 25% of the data is less than and 75% of the data is greater than. Used for finding typical values for skewed data sets.

Third Quartile (Q3): The number that approximately 75% of the data is less than and 25% of the data is greater than. Used for finding typical values for skewed data sets.

Frequency or Sample Size (n)

The frequency or sample size of a data set (n) is not a measure of center, spread or position, but is important bit of information. It tells us how many numbers are in the data set.



Practice Problems Section 1G

1. Answer the following questions:

- Describe a skewed right shape?
- Describe a skewed left shape?
- Define the median average and explain how it is calculated.
- Define the first quartile (Q_1) and explain how it is calculated.
- Define the third quartile (Q_3) and explain how it is calculated.
- Define the interquartile range (IQR) and explain how it is calculated.

2. Answer the following questions:

- If a data set is not normally distributed, what measure of average should we use?
- If a data set is not normally distributed, what measure of typical spread should we use?
- If a data set is not normally distributed, what are the two statistics that typical values are in between?
- If a data set is not normally distributed, approximately what percentage is typical?
- If a data set is not normally distributed, how can we use a box plot to find high outliers in the data set?
- If a data set is not normally distributed, how can we use a box plot to find low outliers in the data set?

(#3-7) Directions: Analyze the following data sets. Go to www.matt-teachout.org, click on the “Statistics” tab, and then the “Data Sets” tab. Open the “Bear” data, the “Health” data, and the “Car” data. Go to www.lock5stat.com and copy and open StatKey. Under the “Descriptive Statistics and Graphs” menu, click on “One Quantitative Variable”. Click on “Edit Data” and copy and paste the indicated data set. Use the graphs and summary statistics to answer the following questions.

3. Bear ages (months)

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- What is the shape of the data set?
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? (*Give the number and the name of the statistic used.*)
- How much typical spread does the data set have?
(*Give the number and the name of the statistic used.*)
- Find two numbers that typical values fall in between.
- List all high outliers in this data set. If there are no high outliers, put “none”.
- List all low outliers in this data set. If there are no high outliers, put “none”.

4. Bear Weights (pounds)

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- What is the shape of the data set?
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? (*Give the number and the name of the statistic used.*)
- How much typical spread does the data set have?
(*Give the number and the name of the statistic used.*)
- Find two numbers that typical values fall in between.



- i) List all high outliers in this data set. If there are no high outliers, put "none".
- j) List all low outliers in this data set. If there are no high outliers, put "none".

5. Women's Systolic Blood Pressure in millimeters of mercury (mm of Hg)

- a) What is the data measuring and what are the units?
- b) How many numbers are in the data set?
- c) What is the shape of the data set?
- d) What is the minimum value?
- e) What is the maximum value?
- f) What is the average (center)? *(Give the number and the name of the statistic used.)*
- g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- h) Find two numbers that typical values fall in between.
- i) List all high outliers in this data set. If there are no high outliers, put "none".
- j) List all low outliers in this data set. If there are no high outliers, put "none".

6. Men's Diastolic Blood Pressure (mm of Hg)

- a) What is the data measuring and what are the units?
- b) How many numbers are in the data set?
- c) What is the shape of the data set?
- d) What is the minimum value?
- e) What is the maximum value?
- f) What is the average (center)? *(Give the number and the name of the statistic used.)*
- g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- h) Find two numbers that typical values fall in between.
- i) List all high outliers in this data set. If there are no high outliers, put "none".
- j) List all low outliers in this data set. If there are no high outliers, put "none".

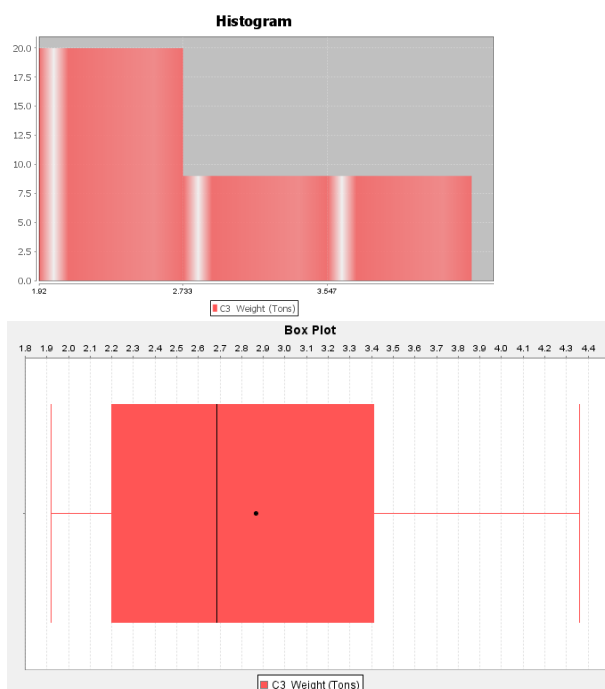
7. Women's Cholesterol in milligrams per deciliter (mg/dL)

- a) What is the data measuring and what are the units?
- b) How many numbers are in the data set?
- c) What is the shape of the data set?
- d) What is the minimum value?
- e) What is the maximum value?
- f) What is the average (center)? *(Give the number and the name of the statistic used.)*
- g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- h) Find two numbers that typical values fall in between.
- i) List all high outliers in this data set. If there are no high outliers, put "none".
- j) List all low outliers in this data set. If there are no high outliers, put "none".



(#8-12) The following graphs and summary statistics were created from the “Car” data at www.matt-teachout.org and Statcato. Use the Statcato graphs and summary statistics to answer the following questions.

8. Weight of various cars in tons.



Descriptive Statistics

Variable	Mean	Standard Deviation
Weight (Tons)	2.864	0.706

Variable	Q1	Median	Q3	IQR
Weight (Tons)	2.198	2.685	3.46	1.262

Variable	Min	Max	Range
Weight (Tons)	1.92	4.36	2.440

Variable	N total
Weight (Tons)	38

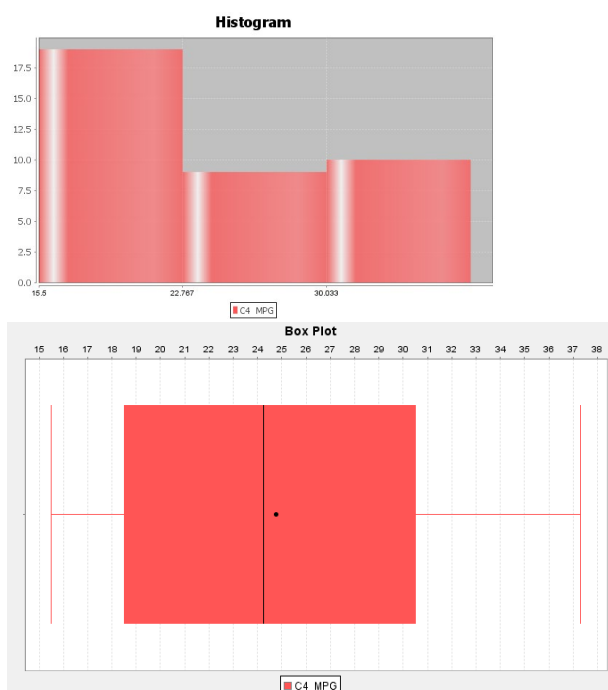
- What is the data measuring and what are the units?
- How many numbers are in the data set?
- What is the shape of the data set?



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

- d) What is the minimum value?
- e) What is the maximum value?
- f) What is the average (center)? *(Give the number and the name of the statistic used.)*
- g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
- h) Find two numbers that typical values fall in between.
- i) List all high outliers in this data set. If there are no high outliers, put "none".
- j) List all low outliers in this data set. If there are no high outliers, put "none".

9. Gas mileage of various cars in miles per gallon (mpg).



Descriptive Statistics

Variable	Mean	Standard Deviation
MPG	24.761	6.547

Variable	Q1	Median	Q3	IQR
MPG	18.425	24.25	30.6	12.175

Variable	Min	Max	Range
MPG	15.5	37.3	21.800

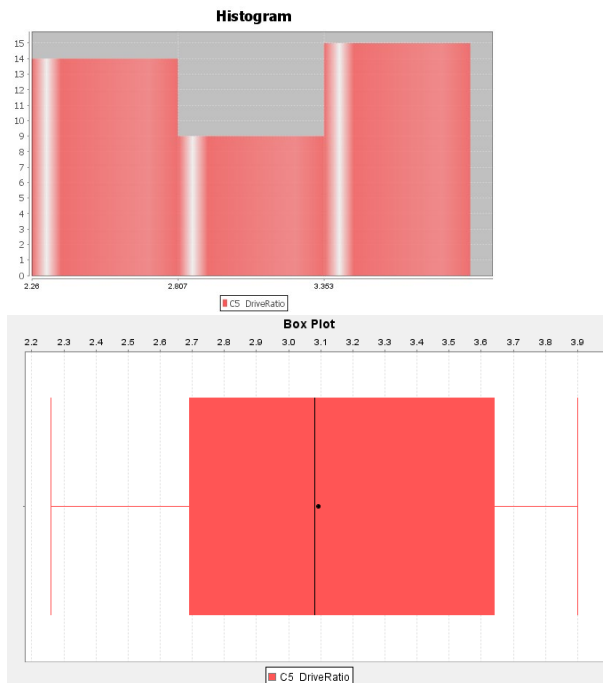
Variable	N total
MPG	38



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

- a) What is the data measuring and what are the units?
- b) How many numbers are in the data set?
- c) What is the shape of the data set?
- d) What is the minimum value?
- e) What is the maximum value?
- f) What is the average (center)? (*Give the number and the name of the statistic used.*)
- g) How much typical spread does the data set have?
(*Give the number and the name of the statistic used.*)
- h) Find two numbers that typical values fall in between.
- i) List all high outliers in this data set. If there are no high outliers, put "none".
- j) List all low outliers in this data set. If there are no high outliers, put "none".

10. The drive ratio of various cars.



Descriptive Statistics

Variable	Mean	Standard Deviation
DriveRatio	3.093	0.518

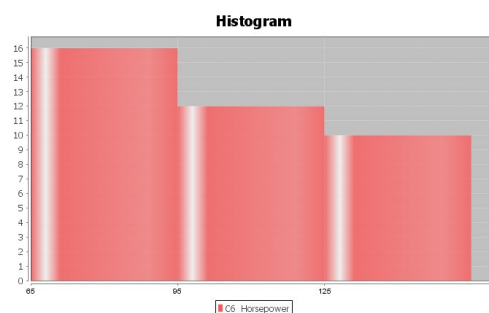
Variable	Q1	Median	Q3	IQR
DriveRatio	2.69	3.08	3.655	0.965

Variable	Min	Max	Range
DriveRatio	2.26	3.9	1.640

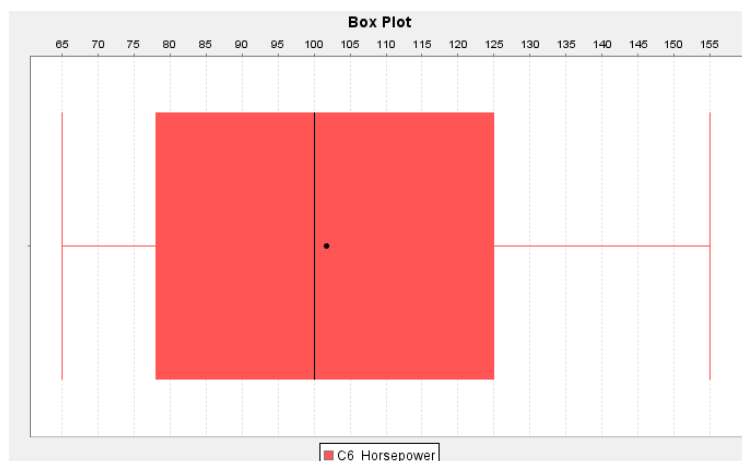
Variable	N total
DriveRatio	38

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- What is the shape of the data set?
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? *(Give the number and the name of the statistic used.)*
- How much typical spread does the data set have? *(Give the number and the name of the statistic used.)*
- Find two numbers that typical values fall in between.
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".

11. The horsepower of various cars.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18



Descriptive Statistics

Variable	Mean	Standard Deviation
Horsepower	101.737	26.445

Variable	Q1	Median	Q3	IQR
Horsepower	77.25	100.0	125.0	47.75

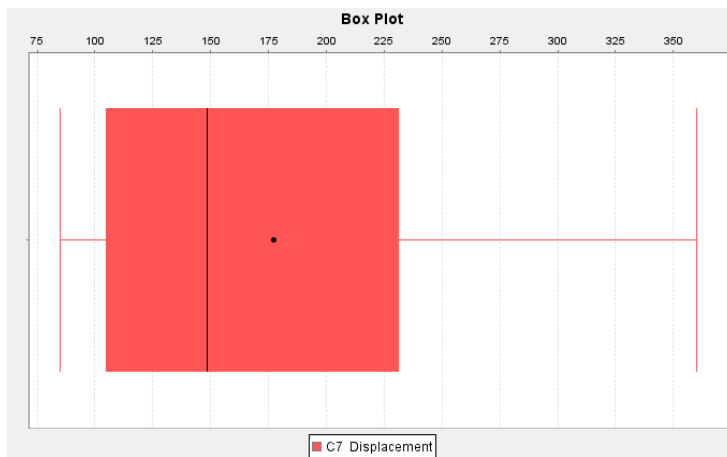
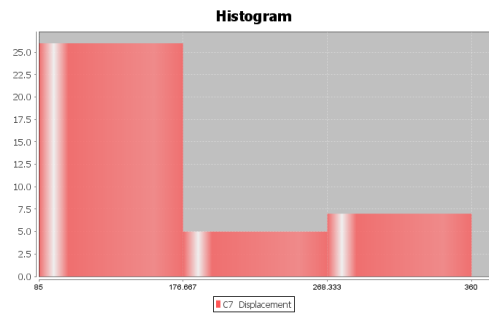
Variable	Min	Max	Range
Horsepower	65.0	155.0	90.0

Variable	N total
Horsepower	38

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- What is the shape of the data set?
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? *(Give the number and the name of the statistic used.)*
- How much typical spread does the data set have? *(Give the number and the name of the statistic used.)*
- Find two numbers that typical values fall in between.
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".



12. The measure of displacement for various cars.



Descriptive Statistics

Variable	Mean	Standard Deviation
Displacement	177.289	88.877

Variable	Q1	Median	Q3	IQR
Displacement	103.25	148.5	237.75	134.5

Variable	Min	Max	Range
Displacement	85.0	360.0	275.0

Variable	N total
Displacement	38

- What is the data measuring and what are the units?
- How many numbers are in the data set?
- What is the shape of the data set?
- What is the minimum value?



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

- e) What is the maximum value?
- f) What is the average (center)? (*Give the number and the name of the statistic used.*)
- g) How much typical spread does the data set have?
(*Give the number and the name of the statistic used.*)
- h) Find two numbers that typical values fall in between.
- i) List all high outliers in this data set. If there are no high outliers, put "none".
- j) List all low outliers in this data set. If there are no high outliers, put "none".

13. Classify each of the following statistics as a measure of center, spread or position.

- a) Q1
- b) Mean
- c) Variance
- d) Standard Deviation
- e) Minimum
- f) Q3
- g) Mode
- h) IQR
- i) Median
- j) Range
- k) Maximum
- l) Midrange

14. Define each of the following statistics and describe when that statistic should be used.

- a) Q1
 - b) Mean
 - c) Variance
 - d) Standard Deviation
 - e) Minimum
 - f) Q3
 - g) Mode
 - h) IQR
 - i) Median
 - j) Range
 - k) Maximum
 - l) Midrange
-



Chapter 1 Review

Key Vocabulary Terms

Data: Information in all forms.

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Random: When everyone in the population has a chance to be included in the sample.

Simple Random Sample: Sample data in which individuals are selected randomly. This method tends to minimize sampling bias and is generally considered a good way to collect data.

Convenience Sample: Sample data that is collected in a way that is easy or convenient. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.

Voluntary Response Sample: Sample data that is collected by putting a survey out into the world and allowing anyone to fill it out. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.

Cluster Sample: Sample data that collects data from groups of people in a population instead of one at a time. The groups should be chosen randomly to avoid sampling bias.

Stratified Sample: Sample data used to compare two or more groups or compare two or more populations. The individuals from each group should be chosen randomly to avoid sampling bias. For example, we may take a random sample of people living in Palmdale, CA and another random sample of people living in Valencia, CA and use the data to compare the average salaries.

Systematic Sample: Sample data that is collected with some type of system like choosing every twentieth person on a list.

Bias: When data does not represent the population.

Sampling Bias: A type of bias that results from collecting data without using a census or random sample. The method of collecting is flawed. For example, using convenience or voluntary response method to collect the data. We can minimize this bias by collecting the data with a census or random sample.

Question Bias: A type of bias that results when someone phrases the question or gives extra information with the goal of tricking the person into answering a certain way. We can minimize this bias by phrasing our questions in a neutral way and not attempt to sway the person giving data.

Response Bias: A type of bias that results when people giving the data do not answer truthfully or accurately. To minimize this bias, we should collect the data anonymously and assure the person giving the data that the data will be used for scientific purposes and will not be released.

Non-response Bias: A type of bias that results when people refuse to participate or give data. To minimize non-response bias, you may give an incentive like a gift card to encourage people to give data.



Deliberate Bias: A type of bias that results when the people collecting the data falsify the reports, delete data, or decide to not collect data from certain groups in the population. To minimize deliberate bias, the people collecting and analyzing the data need to have good ethics. They should not falsify reports, delete data or leave out groups from the population.

Experimental Design: A scientific method for controlling confounding variables and proving cause and effect.

Observational Study: Collecting data without controlling confounding variables. This type of data cannot prove cause and effect.

Explanatory Variable: The independent or treatment variable. In a cause and effect experiment, this is the cause variable.

Response Variable: The dependent variable. In a cause and effect experiment, this is the variable that measures the effect.

Treatment Group: The group of people or objects that has the explanatory variable. In an experiment involving medicine, this would be the group that receives the medicine.

Control Group: The group of people or objects that is used to compare and does not have the explanatory variable. In an experiment involving medicine, this would be the group that receives the placebo.

Confounding Variables: Also called lurking variables. Other variables that might influence the response variable other than the explanatory variable being studied.

Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

Placebo: A fake medicine or fake treatment used to control the placebo effect.

Percentage: A statistic calculated from categorical data that measures the part out of 100.

Proportion: The decimal equivalent to a percentage.

Sample Size (n): Also called the sample frequency or sample count. This is the number of people or objects represented in the sample data. If we collected data from 35 people, then the sample size would be $n = 35$.

Mean: A measure of center or average for quantitative data that balances the distances. The mean average is only accurate if the quantitative data is normal (bell shaped). Hence, the mean average is the center or average used for normal quantitative data.

Median: A measure of center or average for quantitative data that is found by finding the center of the data when the data values are in order. The median is the most accurate center or average when the data is skewed left, skewed right, or not normal.

Mode: The number or numbers that appear most often in a quantitative data set. The mode is used as a measure of center or average.

Midrange: A quick measure of center or average that is half way between the min and max of a quantitative data set. It is generally not very accurate, but easy to calculate.

Standard Deviation: The most accurate measure of typical spread for normal (bell shaped) quantitative data. The standard deviation measures how far typical values are from the mean on average. The standard deviation is only accurate if the quantitative data is normal (bell shaped).

Variance: A measure of spread used in ANOVA testing. The variance is the square of the standard deviation and is only accurate when data is normal (bell shaped).



Interquartile Range (IQR): The most accurate measure of spread for skewed or non-normal data. The IQR measures how far typical values are from each other in skewed or non-normal data. IQR is calculated by subtracting the Third Quartile (Q3) minus the First Quartile (Q1).

Range: A quick measure of spread that measures the distance between the max and min of a quantitative data set. It is easy to calculate (Max – Min), but is not an accurate measure of typical spread, since it does not involve typical values in the data.

First Quartile (Q1): The divider that approximately 25% of the quantitative data values are less than. Q1 is the bottom range of typical values for skewed or non-normal data. Typical values are between Q1 and Q3 in skewed or non-normal data. Q1 is considered a measure of position.

Third Quartile (Q3): The divider that approximately 75% of the quantitative data values are less than. Q3 is the top range of typical values for skewed or non-normal data. Typical values are between Q1 and Q3 in skewed or non-normal data. Q3 is considered a measure of position.

Max: The largest number in a quantitative data set. Considered a measure of position.

Min: The smallest number in a quantitative data set. Considered a measure of position.

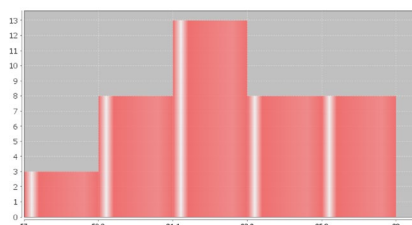
Categorical Data Analysis

- **To convert a decimal proportion into a percentage => Multiply by 100 and put on the % symbol.** (This will move the decimal two places to the right.)
- **To convert a percentage into a decimal proportion => Remove the % symbol and Divide by 100.** (This will move the decimal two places to the left.)
- **To calculate the proportion for each categorical variable:** $\text{Proportion} = \frac{x}{n} = \frac{\text{Amount (\# of successes)}}{\text{Total Frequency (Sample Size)}}$
(StatKey calculate counts and proportions for you.)
- **Round Proportions to the thousandths place.** (Three numbers to the right of decimal point. StatKey round to the thousandths place.)

Quantitative Data Analysis

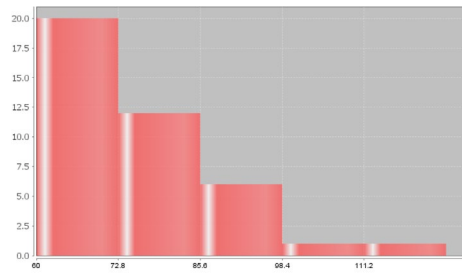
Shapes

Normal (Bell Shaped, Unimodal and Symmetric)

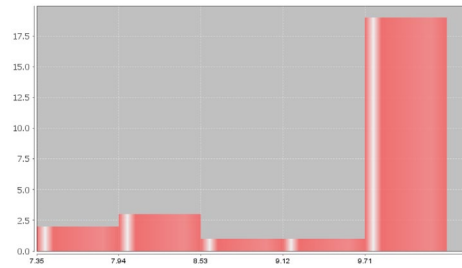


This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

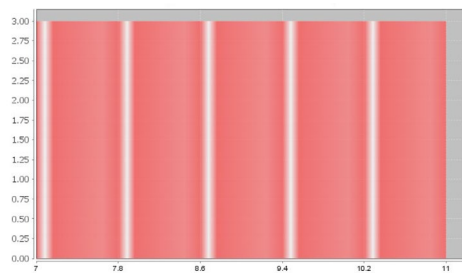
Skewed Right (Positively Skewed)



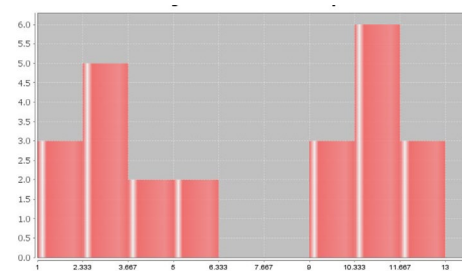
Skewed Left (Negatively Skewed)



Uniform



Bimodal



Shape determine what statistics are accurate!



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Normal Quantitative Data

Center (Average): Mean

Typical Spread: Standard Deviation

Typical Values Between: Mean – Standard Deviation and Mean + Standard Deviation
(One Standard Deviation above and below the mean, Middle 68% of the data)

High Outliers: Data Values \geq Mean + (2 x Standard Deviation)
(Two standard deviations above the mean, Top 2.5% of the data)

Low Outliers: Data Values \leq Mean – (2 x Standard Deviation)
(Two standard deviations below the mean, Bottom 2.5% of the data)

Skewed or Non-normal Quantitative Data

Center (Average): Median

Typical Spread: Interquartile Range (IQR)

Typical Values Between: 1st quartile (Q1) and 3rd quartile (Q3)
(Middle 50% of data values)

High Outliers: Data Values \geq Q3 + (1.5 x IQR) Automatically calculated in Box-Plot!

Low Outliers: Data Values \leq Q1 – (1.5 x IQR) Automatically calculated in Box-Plot!

Chapter 1 Review Problems

1. Tell if the following data is categorical or quantitative and explain why.

- The types of cars in the different parking lots.
- The average number of hours spent practicing ping-pong.
- Areas in North Dakota that have wild mustangs.
- Each person is asked if he or she wear glasses, contacts, neither, or both.
- The average speed of racecars at the Indianapolis 500.
- Exam scores for various students on a history exam.

2. Jim wants to know how much money the average working COC student makes. Describe how Jim could use each of the following techniques to collect data. For each technique, will there be a significant amount of sampling bias or not too much sampling bias?

- Systematic
- Voluntary Response
- Random Sample
- Convenience Sample
- Cluster Sample



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

- f) Stratified Sample
- g) Simple Random Sample
- h) Census

3. Define the following key terms and give an example of each.

- a) Population
- b) Census
- c) Sample
- d) Random
- e) Bias
- f) Statistic

4. Describe and give an example of each of the following types of bias. Also state how a person collecting and analyzing data, can avoid these biases.

- a) Sampling Bias
- b) Question Bias
- c) Response Bias
- d) Deliberate Bias
- e) Non-Response Bias

5. Rachael needs to do an experiment that will show that wearing nicotine patches cause a person to stop smoking. Set up the experiment for Rachael. What is the explanatory variable? What is the response variable? Write a description of the experiment and include the following. What are some confounding variables that she will need to control? How can Rachael control the confounding variables? Include a description of how Rachael use a double blind placebo to control the placebo effect. Describe the treatment group and the control group in the experiment.

6. Compare and contrast the similarities and differences between an experiment and an observational study. How can we tell if we should use an experiment or an observational study?

7. Explain the following.

- a) Explain how to round a decimal to a given place value.
- b) Explain how to convert a decimal proportion into a percentage.
- c) Explain how to convert a percentage into a decimal proportion.
- d) Explain how to calculate a percentage by using an amount and a total from categorical data.
- e) Explain how to calculate an estimated amount by using a percentage and a total from categorical data.

8. Convert the following proportions into percentages. Do not round your answer.

- a) 0.0722
- b) 0.0041
- c) 0.563
- d) 0.0005



9. Convert the following percentages into decimal proportions. Do not round your answer.
- 35.9%
 - 4.823%
 - 0.026%
 - 0.389%
10. A company has 74 employees. Of those employees 11 are managers, 27 are full-time employees and 36 are part-time employees. Use this information to answer the following questions.
- What proportion of the employees are managers? *(Round your answer to the thousandths place.)*
 - What percentage of the employees are managers? *(Round your answer to the tenths place.)*
 - What proportion of the employees are full-time employees?
(Round your answer to the thousandths place.)
 - What percentage of the employees are full-time employees? *(Round your answer to the tenths place.)*
 - What proportion of the employees are part-time employees?
(Round your answer to the thousandths place.)
 - What percentage of the employees are part-time employees? *(Round your answer to the tenths place.)*
 - Calculate the percent of increase between managers and full-time employees. Is there a significant difference between the percentages? Explain why.
 - Calculate the percent of increase between full-time and part-time employees. Is there a significant difference between the percentages? Explain why.
11. According to an online article, approximately 60% of the voting population in the U.S. votes during a presidential election year. According to a census, there are approximately 41,743 people living in Saugus, CA. If 60% of them vote in the next presidential election, how many people do we expect to vote in Saugus?
12. Describe and draw a histogram for each of the following shapes.
- Normal
 - Skewed Right
 - Skewed Left
 - Uniform
 - Bimodal
13. Classify each of the following quantitative statistics as a measure of center, spread or position. Also, describe when that statistic should be used.
- Q1
 - Mean
 - Variance
 - Standard Deviation
 - Minimum
 - Q3
 - Mode
 - IQR
 - Median
 - Range
 - Maximum
 - Midrange
14. Answer each of the following questions about quantitative data analysis.
- What measure of center (average) should we use if the data is normal?
 - What measure of center (average) should we use if the data is not normal?
 - What measure of spread (variability) should we use if the data is normal?
 - What measure of spread (variability) should we use if the data is not normal?

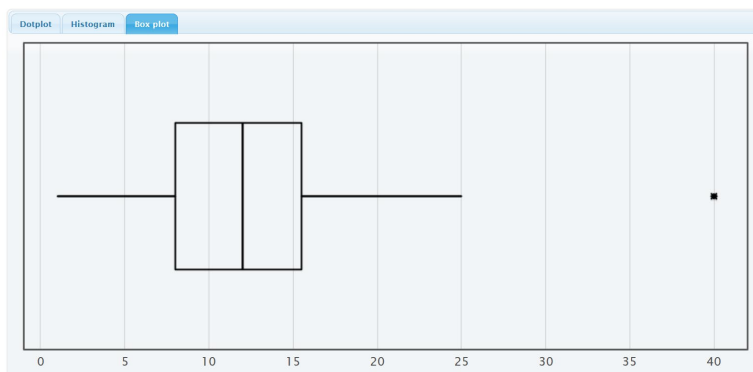
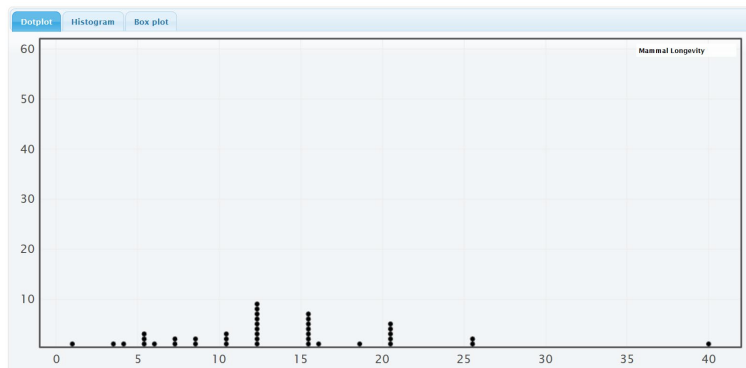


- e) How do we find two numbers that typical values fall in between if the data is normal?
- f) How do we find two numbers that typical values fall in between if the data is not normal?
- g) What is the formula for the high outlier cutoff if the data is normal?
- h) What is the formula for the high outlier cutoff if the data is not normal?
- i) What is the formula for the low outlier cutoff if the data is normal?
- j) What is the formula for the low outlier cutoff if the data is not normal?
- k) How do we determine if a data value is an outlier when the data is normal?
- l) How do we determine if a data value is an outlier when the data is not normal?

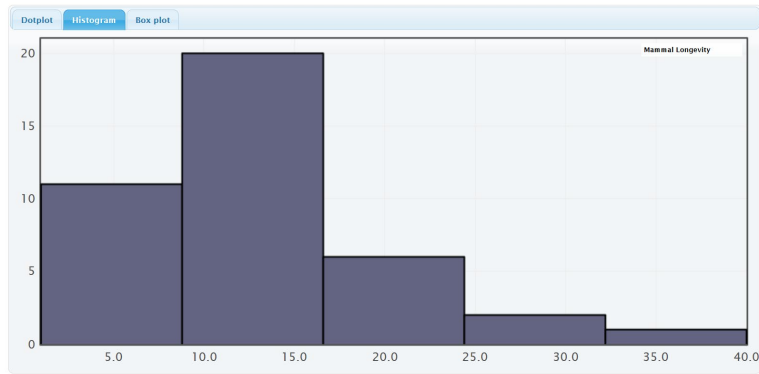
15. The following graphs and statistics were calculated with StatKey and describe the number of years mammals live. Use the graphs and statistics to answer the following questions.

Summary Statistics

Statistic	Value
Sample Size	40
Mean	13.150
Standard Deviation	7.245
Minimum	1
Q_1	8.000
Median	12.000
Q_3	15.500
Maximum	40



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18



- What is the data measuring and what are the units?
- How many mammals are in the data set?
- What is the shape of the data set?
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? (*Give the number and the name of the statistic used.*)
- How much typical spread does the data set have? (*Give the number and the name of the statistic used.*)
- Find two numbers that typical values fall in between.
- List all high outliers in this data set. If there are no high outliers, put "none".
- List all low outliers in this data set. If there are no high outliers, put "none".

16. The following graphs and statistics were calculated with Statcato and describe the number of years employees have been employed at a company. Use the graphs and statistics to answer the following questions.

Descriptive Statistics

Variable	Mean	Standard Deviation
years employed	7.345	1.376

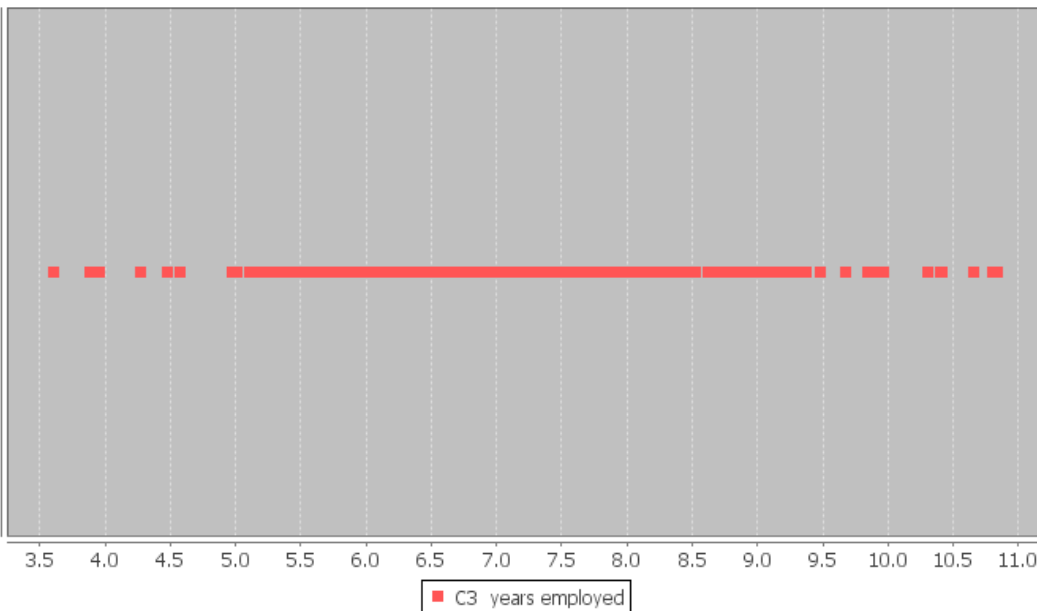
Variable	Q1	Median	Q3	IQR
years employed	6.4	7.35	8.3	1.9

Variable	Min	Max	Range
years employed	3.6	10.8	7.2

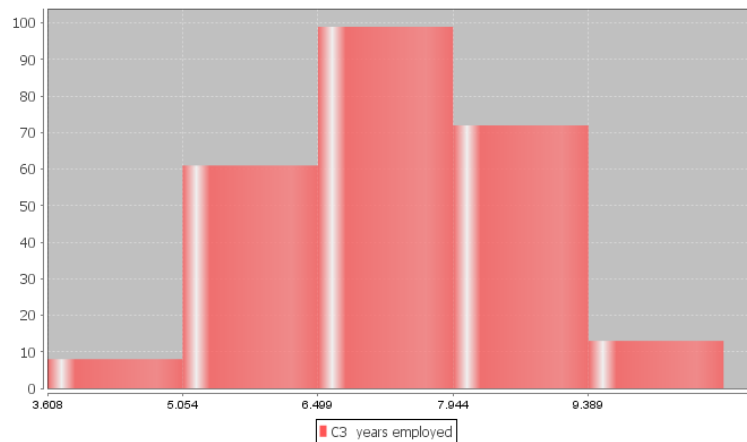
Variable	N total
years employ	253



Dot Plot



Histogram



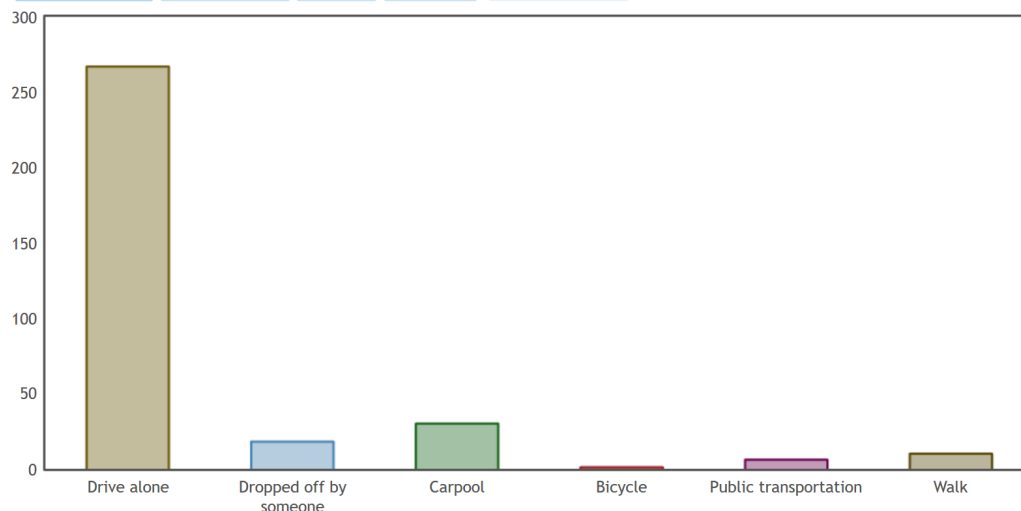
- What is the data measuring and what are the units?
- How many employees are in the data set?
- What is the shape of the data set?
- What is the minimum value?
- What is the maximum value?
- What is the average (center)? *(Give the number and the name of the statistic used.)*
- How much typical spread does the data set have? *(Give the number and the name of the statistic used.)*
- Find two numbers that typical values fall in between.
- Calculate the high-outlier cutoff. Give approximate values of the high outliers in this data set. If there are no high outliers, put "none".
- Calculate the low-outlier cutoff. Give approximate values of the low outliers in this data set. If there are no low outliers, put "none".



17. Statistics students were asked what mode of transportation they take to get to school. Use the following bar chart and statistics to answer the following.

StatKey Descriptive Statistics for One Categorical Variable

Custom Dataset Show Data Table Edit Data Upload File Change Column(s)



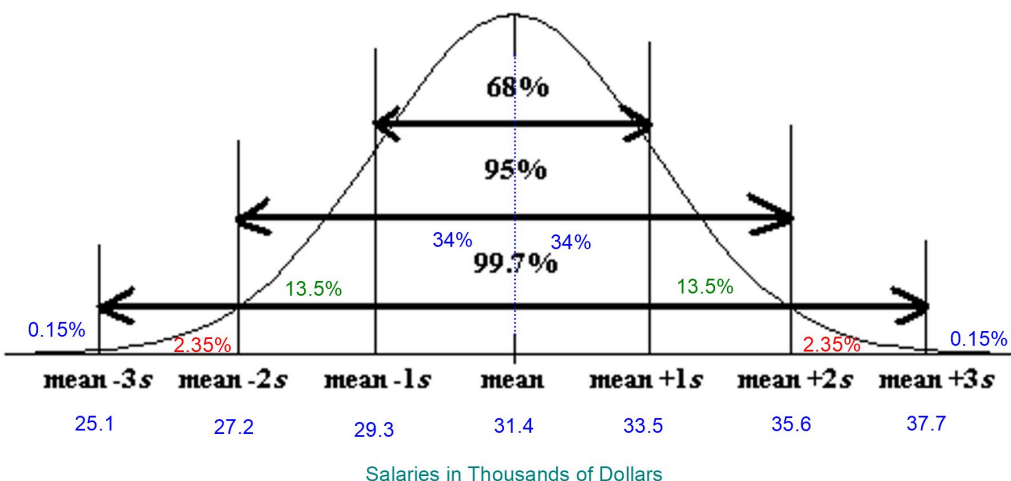
Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

- What was the most population mode of transportation?
- What was the least population mode of transportation?
- How many statistics students walked to school?
- What proportion of statistics students were dropped off by someone? Do not calculate the answer. Use the table provided. Do not round the answer.
- What percentage of the statistics students use public transportation? Do not calculate the answer. Use the table provided and convert the answer into a percentage. Do not round the answer.



18. The salaries of employees at a company are normally distributed with a mean of 31.4 thousand dollars and a standard deviation of 2.1 thousand dollars. Use the Empirical Rule graph below to answer the following questions about this normal quantitative data.



- What percent of the salaries are between 29.3 thousand dollars and 31.4 thousand dollars?
- What percent of the salaries are 33.5 thousand dollars or more?
- Typical salaries are between what two values?
- What is the high outlier cutoff?
- What is the low outlier cutoff?
- What is the average salary for employees of this company?

Chapter 2: Estimating Population Parameters

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Sampling Distribution: Take many random samples from a population, calculate a sample statistic like a mean or percent from each sample and graph all of the sample statistics on the same graph.
The center of the sampling distribution is a good estimate of the population parameter.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Point Estimate: When someone takes a sample statistic and then claims that it is the population parameter.

Margin of Error: Total distance that a sample statistic might be from the population parameter. For normal sampling distributions and a 95% confidence interval, the margin of error is approximately twice as large as the standard error.

Standard Error: The standard deviation of a sampling distribution. The distance that typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distributions is usually close to the population parameter, the standard error tells us how far typical sample statistics are from the population parameter.

Confidence Interval: Two numbers that we think a population parameter is in between. Can be calculated by either a bootstrap distribution or by adding and subtracting the sample statistic and the margin of error.

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval. The center of the bootstrap distribution is the original real sample statistic.

Introduction: The goal of learning Statistics or Data Science is to be able to analyze data to learn about populations in the world around us. The best way to understand a population is collect and analyze unbiased data from that population, namely a census. The trouble is we rarely have an unbiased census. It is sometimes impossible to collect data from everyone in a population. We have to rely on samples, small subgroups of the population. The next few chapters deal with the subject of using samples to understand populations. This is sometimes called "inferential statistics". We will start by trying to distinguishing between population parameters from sample statistics.

Section 2A – Statistics and Parameters

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not represent the population.

The goal of collecting and analyzing data is to understand the world around us. To this end, our goal is understand populations. The population is all of the people or objects you plan to study. A population can be large (like all



*This chapter is from **Introduction to Statistics for Community College Students**, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

people living in Brazil) or small (like all students in a particular statistics class). It goes without saying that the larger the population the more difficult it is to understand.

The best data for representing populations is an unbiased census. A census is an attempt to collect data from everyone in a population. A census is easier if we have a small population like the people in a particular statistics class. The advantage of collecting an unbiased census is that we can calculate population values (parameters) directly with reasonable certainty. Governments may sometimes attempt to do a census and collect data on all of the people living in a particular country. It should be noted that though they attempt to get data on everyone, they rarely succeed. There will always be some people fall through the cracks and are not represented in the census. An unbiased census of a large population still represents a high percentage of the people, so is generally better than a small sample of people.

A data scientist rarely has the ability to collect a census unless the population is relatively small. People that work in statistics and data science usually rely on collecting samples. Remember a sample is a small subgroup of the population. It is usually less than 10% of the population and is often significantly less than 10%. If the sample is unbiased, we then try to analyze the sample data and make guesses as to what is happening at the population level. Therefore, a data scientist or statistician needs to be able to use sample values (statistics) to figure out approximate population value (parameters).

Statistic: A number calculated from sample data in order to understand the characteristics of the data.

Parameter: A population value. It can be calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be.

It is very important to note that statistics and parameters are not the same thing. A statistic calculated from 250 people in a sample will often be very different from the actual population parameter from millions of people. The question that is important to ask is how far off is the sample statistic from the population parameter? That is sometimes called “margin of error” and is a key topic in this chapter.

Common Statistics

\bar{x} : (“x-bar”) Sample mean average

s : Sample standard deviation (typical distance from the sample mean)

s^2 : Sample variance (sample standard deviation squared)

\hat{p} : (“p-hat”) Sample proportion (sample percentage)

n : Sample size or frequency (number of people or objects in the sample)

r : Sample correlation coefficient (measures quantitative relationships between samples)

b_1 : Sample slope (The slope of a regression line calculated from sample data.)

b_0 : Sample Y-intercept (The Y-intercept of a regression line calculated from sample data.)

Common Parameters

μ : (“mu”) Population mean average

σ : (“sigma”) Population standard deviation (typical distance from the population mean)

σ^2 : Population variance (population standard deviation squared)

π : (“pi”) Population proportion (population percentage) (*Some people use “p” for population proportion.*)

N : Population size or frequency (number of people or objects in the population)



ρ : (“rho”) Population correlation coefficient (measures quantitative relationships between populations.

Note this is not a “p”. It is the Greek letter “rho”).

β_1 : Population slope (The slope of the population regression line. Used when studying quantitative relationships between populations.)

β_0 : Population Y-intercept (The Y-intercept of the population regression line. Used when studying quantitative relationships between populations.)

Let us look at some examples of using statistics and parameters. It is important to be able to identify if a number used is a statistic or a parameter and what letter we might use in the computer program.

Example

“We think the mean average ACT score for all high school students is about 22. The mean average ACT score for a random sample of 85 high school students was 21.493”

$\mu = 22$ (parameter)

$n = 85$ (statistic)

$\bar{x} = 21.493$ (statistic)

Example

“A random sample showed that 13.2% of adults were infected, but this indicates that the population percentage could be 17%”. (Note: Computer programs often require you to convert the percentages into decimal proportions.)

$\hat{p} = 0.132$ (statistic)

$\pi = 0.17$ (parameter)

Example

The standard deviation for the heights of all women is thought to be about 2.5 inches. A random sample of women heights had a standard deviation of 2.618 inches.

$\sigma = 2.5$ (parameter)

$s = 2.618$ (statistic)

Example

“Sample data indicated that the correlation coefficient was 0.239 and the slope was 47.3 dollars per pound. Let’s compare these to the population claims that the correlation coefficient is zero and the slope is about 50 dollars per pound.”

$r = 0.239$ (statistic)

$b_1 = 47.3$ (statistic)

$\rho = 0$ (parameter)

$\beta_1 = 50$ (parameter)



Problem Set Section 2A

1. Describe each of the following symbols. What does the symbol represent? Is the symbol describing a sample statistic or a population parameter?

$N, n, \pi, \hat{p}, \mu, \bar{x}, \sigma, s, \sigma^2, s^2, \rho, r, \beta_1, b_1$

(#2-25) *Directions: Determine if the numbers in the following clips from magazines and newspapers are describing a population parameter or a sample statistic. In each case, give the symbol we would use for the parameter or statistic.* ($N, n, \pi, \hat{p}, \mu, \bar{x}, \sigma, s, \sigma^2, s^2, \rho, r, \beta_1, b_1$)

2. "Our study found that of the 200 people tested in the sample, only 3% showed side effects to the medication."
3. "It has been speculated for years that the mean average height of all men is 69.2 inches, but our sample data disagrees with this. Our sample mean average was 69.5 inches."
4. "The standard deviation for all humans is about 1.8 degrees Fahrenheit. A random sample of 52 people found a standard deviation of 1.739 degrees Fahrenheit".
5. "We tested a sample of 300 incoming college freshman and found that their mean average IQ was 101.9 with a standard deviation of 14.8".
6. "The mean average human body temperature has long been thought to be 98.6 degrees Fahrenheit, but our sample of 63 randomly selected adults had a mean average was 98.08".
7. "The mean average number of units that students take per semester is about 12, but when we took a random sample of 160 college students found that the mean average was 12.37 units."
8. "A public opinion poll showed that 47.2% of voters would vote for the candidate, but when the votes or entire population were counted we found that only 41.3% voted for the candidate."
9. "According to the California Department of Finance, the Los Angeles county population as of January 2015 was approximately 10,136,559 people."
10. "We want to check and see if the population correlation coefficient could be zero. The sample correlation coefficient was 0.338."
11. "Many experts think that the population slope for weight gain in these type of bears is about 3 pounds per month, but the sample slope from 54 bears was 2.7055 pounds per month."
12. "A random sample of 40 men found that the sample variance for systolic blood pressure was 109.474, but this indicates that the population variance could be as high as 173."
13. "According to the 2015 U.S. census, approximately 78% of U.S. households own a computer. A random sample of 165 households found that 81.2% of them owned a computer."
14. "We think that the population correlation coefficient is zero. The sample correlation coefficient was 0.0371."
15. "IQ tests are supposed to have a population mean of 100 and a population standard deviation of 15 IQ points. This could be correct since our random sample data had a mean of 97.7 and the standard deviation of 15.3 IQ points."
16. "When analyzing the relationship between the amount of mercury and the pH of Florida lakes, we found a sample slope of -0.152. We are wondering if the population slope could be zero."
17. "We believe that the population mean average pH of Florida lakes is approximately 6.7, yet our sample data from 53 randomly selected lakes had a mean of 6.591."



18. "While the sample variance is 37.882, we think the population variance could be as high as 50."
 19. "We believe there are approximately 59,530 people currently living in Canyon Country, CA."
 20. "A random sample of 60 adults found that 21.7% of them had this characteristic. However, we think the population percentage is probably closer to 15%."
 21. "The mean average weight of the 10 male lions was 437.2 pounds. Most people believe that the mean average weight of all male lions is closer to 420 pounds."
 22. "The correlation coefficient for the ordered pair sample data was 0.922. This seems very significant, but does this indicate that the population correlation coefficient is 1?"
 23. "We analyzed the gas usage and distance for large 18-wheeler trucks and found the sample slope to be 6.23 miles per gallon. Articles online indicate that the population slope for all 18-wheeler trucks is closer to 5.9 miles per gallon."
 24. "The sample standard deviation was approximately \$3.78. We want to see if the population standard deviation could be \$3.50."
 25. "A random sample of 38 cars, found that the mean average displacement was 177.289 and the standard deviation was 88.877."
-

Section 2B – Sampling Variability and Sampling Distributions

If you wanted to study baseball players, would you only study one baseball player? If you wanted to study bears, would you only study one bear? The answer of course is no. When studying a topic like bears or baseball players, we should look at many different bears, many different baseball players. The problem with studying samples is that we usually only collect one sample at a time. We cannot learn about the behavior and variability in samples if we only look at one sample. We need to look at hundreds or even thousands of samples.

Sampling Distributions

Suppose we take many, many random samples from a population. From each random sample, we calculate a statistic like the sample mean average. If we put all of those sample means on the same graph, we have created a "sampling distribution". Sampling distributions are one of the best ways to understand random samples and sampling variability.

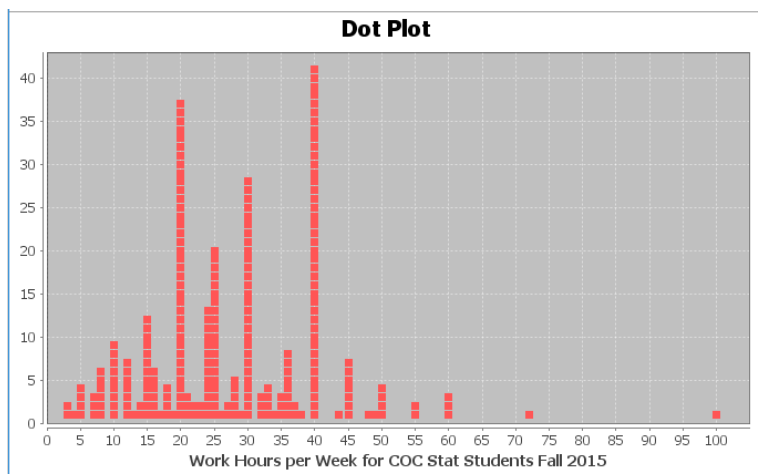
In the real world, a data scientist has only one random sample and may have no idea what a population parameter is. In this example, we will be creating a sampling distribution by take random samples from a census. We will assume the census is unbiased. With an unbiased census, we will know what the population parameter is. That way we can compare our sample statistics to the parameter and study the variability.

Example: Work Hours per Week for working COC Statistics Students (Fall 2015 semester)

We will start by looking at a census of the work hours of all of the working Math 140 students in the fall 2015 semester. It should be noted that we are only studying the statistics students that said they work in addition to going to school. We removed all of the students that work zero hours. We will take many random samples of size 50 from this census data and create a sampling distribution for various statistics.

Census Data (Work Hours per Week for working COC Stat Students Fall 2015)





Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0

We see that the census data is skewed right with a population mean average of 27.283 hours per week, a population standard deviation of 12.969 hours per week, and a population median of 25 hours per week. We will assume that the census was unbiased and these are parameters.

Population mean = 27.283 hours per week
 Population standard deviation = 12.969 hours per week
 Population median = 25 hours per week

We learned in chapter 1 that random samples tend to minimize sampling bias, so are better representations of the populations than other samples that are not random. Does this mean that random samples are perfect representations of the population? Let us see.

Sample 1: Here is one random sample of 50 statistics students from the work hours census data.

Descriptive Statistics

Variable	Mean	Standard Deviation
work hours random sample1	26.93	11.266

Variable	Median
work hours random sample1	24.0

Variable	Sample Size
work hours random sample1	50



We see that the sample mean was 26.93 hours per week, the sample standard deviation was 11.266 hours per week, and the sample median was 24 hours per week. Notice that all of these sample statistics are different from the population parameters.

Sample 1 mean = 26.93 hours per week

Population mean = 27.283 hours per week

Sample 1 standard deviation = 11.266 hours per week

Population standard deviation = 12.969 hours per week

Sample 1 median = 24 hours per week

Population median = 25 hours per week

Sample 2: Let us take another random sample of 50 statistics students work hours from the population.

Descriptive Statistics

Variable	Mean	Standard Deviation
work hours random sample2	29.5	12.732

Variable	Median
work hours random sample2	30.0

Variable	Sample Size
work hours random sample2	50

We see that the sample mean was 29.5 hours per week, the sample standard deviation was 12.732 hours per week, and the sample median was 30 hours per week. Notice these sample statistics are also different from the population parameters. They are also different from the last random sample.

Sample 2 mean = 29.5 hours per week

Sample 1 mean = 26.93 hours per week

Population mean = 27.283 hours per week

Sample 2 standard deviation = 12.732 hours per week

Sample 1 standard deviation = 11.266 hours per week

Population standard deviation = 12.969 hours per week

Sample 2 median = 30 hours per week

Sample 1 median = 24 hours per week

Population median = 25 hours per week

These examples show us that random sample statistics will usually be different from the population parameters. Random sample statistics will also be different from each other. Every time we take another random sample from the same population, we will get different values. This is the principle of “sampling variability” and is a major roadblock on the quest to estimating population parameters.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.



Let us continue taking random samples from the population of working statistics students in fall 2015. Every time we take a random sample, we keep getting different values and different statistics. Hardly any of the samples are close to the population parameter. In this example, we will focus on the mean. Remember the population mean average was 27.283 hours per week. No matter how many random samples we take, the sample means are usually different from the population mean of 27.283 hours per week. Every sample has a “margin of error”.

Margin of Error: How far off a sample statistic can be from the population parameter.

In the first random sample, the sample mean was 26.93 hours per week. So the sample mean of 26.93 hours per week was 0.353 hours lower than the population mean of 27.283 hours per week. This is the margin of error.

In the second random sample, the sample mean was 29.5 hours per week. So the sample mean of 29.5 hours per week was 2.217 hours higher than the population mean of 27.283 hours per week. Again, that is the margin of error for that sample.

What does this tell us?

The principle of sampling variability tells us that sample statistics will usually be off from the population parameter. In other words, almost all samples have a margin of error. Sometimes random samples are closer to the population parameter like sample 1 and sometimes the random samples are farther away like sample 2.

Important Note: If you know the population parameter, then it is relatively easy to calculate the margin of error (sample statistic – population parameter). Most of the time, we are working with sample data, so have no idea what the population parameter is. In that case, it is much more difficult to figure out the potential margin of error. Formulas were developed in order to estimate what the margin of error could be.

Point Estimates

People are usually very interested to know population values. However, we rarely ever know the population parameter. In the real world, we usually only have one random sample. Sometimes, a person will simply tell you that the sample statistic is the population parameter. This is called a “point estimate” and tends to create a lot of confusion for people.

Point Estimate: When someone takes a sample statistic and then claims that it is the population parameter.

In an article published by a health website, the author states that the population average weight of all men in America is 196 pounds. As with most articles, this is a guess about the population average and is not the actual population average weight of men. We call this a “point estimate”. Someone took a sample of men and weighed them. We do not know the sample size or if the sample was even random. They calculated the sample average and found it to be 196 pounds. Since no one really knows the population average weight of all men in the U.S., the author simply tells us the sample average is the population average.

Think about the principle of sampling variability that we just learned. We said that a sample statistic usually has a margin of error is off from the population parameter. Yet people reading the article believe that the population average weight of all men in the U.S. is exactly 196 pounds.

Population parameters may be calculated if we had an unbiased census, but remember that is rare. (Certainly, we do not have an unbiased census of the weights of all men in the U.S.) Usually, we have one random sample. When reading an article that claims to know a population parameter like a population mean or population percentage, it is important to realize that it is just a guess about the population parameter, and that guess probably came from a sample. Sample statistics can be very off from the actual population parameter.



Sampling Distributions for Sample Mean Averages

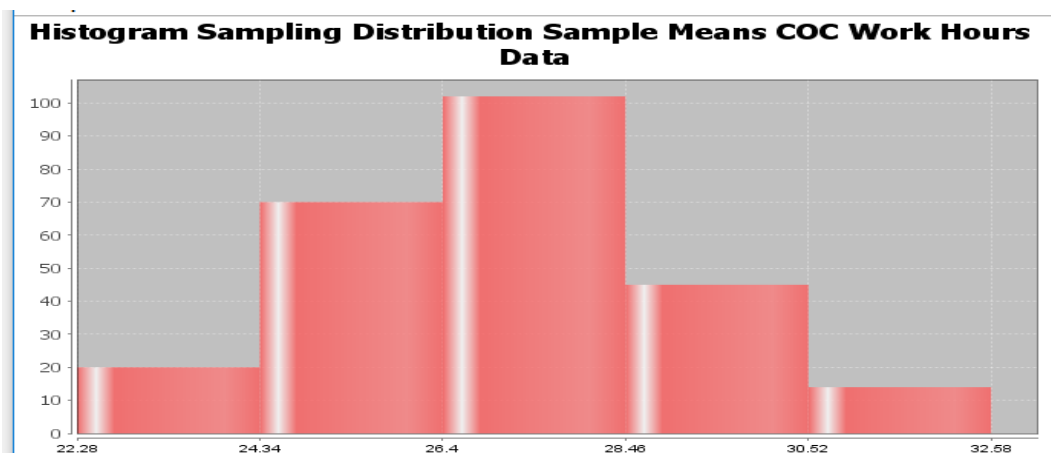
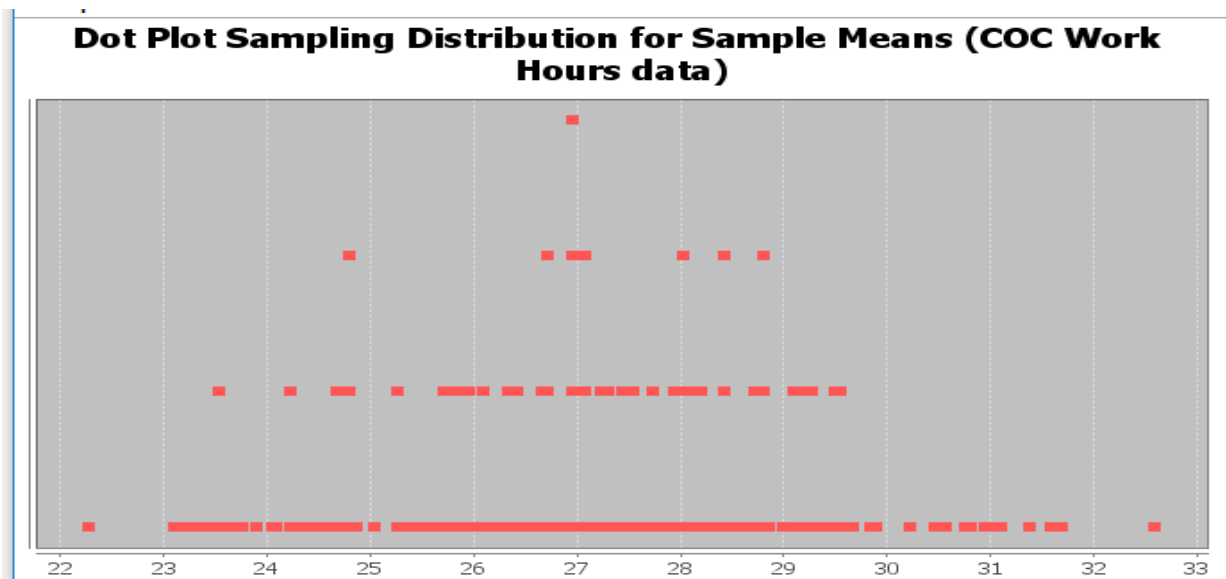
Let us go back to the example of working COC statistics students in the fall 2015 semester. We have seen that the population mean average is 27.283 hours per week, but the two random samples of 50 statistics students gave sample means that have both been off from that population mean.

Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0

Let us continue to collect random samples of size 50 and calculate sample means. We collected 251 random samples and calculated 251 sample means. If we put all of the sample means on the same graph, we can create a sampling distribution.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Here is the sampling distribution we created with Statcato. Each dot in the sampling distribution represents the sample mean of a random sample. We also created a histogram of the sampling distribution to better judge the shape. Notice a few things.

Descriptive Statistics

Variable	Center (Mean) of Sampling Distribution	Standard Error
Sampling Distribution for Sample Means (COC stat students work hours)	27.127	1.916

Variable	Min	Max
Sampling Distribution for Sample Means (COC stat students work hours)	22.28	32.58

Variable	Total Number of Random Samples
C3 Sampling Distribution for Sample Means (COC stat students work hours)	251

- We took 251 random samples and calculated 251 sample means. We see sampling variability in action. The population mean is 27.283 hours per week but sample means ranged between 22.28 hours and 32.58 hours. Random sample means are usually not the same as each other and can be very different from the population mean.
- Despite the population being skewed right, the sampling distribution for these sample means is normal. This is often referred to as the “Central Limit Theorem”.
- The center of the sampling distribution is 27.127 hours. This is not the mean of a sample. It is the mean average of all the sample means. Notice that the center of the sampling distribution is very close to the population mean of 27.283 hours.
- We also calculated the “standard error”. This is the standard deviation of the sampling distribution (or the standard deviation of all the sample statistics) and is an important measure of sampling variability. Think of it this way. The standard error tells us how far typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distribution is 27.127 hours and is pretty close to the population parameter of 27.283 hours, the standard error tells us how far typical sample statistics are from the population parameter. In this case, it tells us that typical sample means are approximately within 1.916 hours of the population mean.

Important Note: Do not confuse the standard error with the margin of error. The standard error tells us how far typical sample statistics are from the population value, but not all random samples are typical. Remember we learned from the empirical rule that typical for normal data represents only the values that are within one standard deviation from the mean (middle 68%). Usually sample values can be up to two standard deviations from the mean (middle 95%). So early statisticians thought that the margin of error should be about twice as large as the standard error. This is still a common formula for margin of error.

Margin of Error = $2 \times$ Standard Error

Sampling Distributions for Sample Standard Deviations

In data science, we often want to estimate many different population parameters besides the mean average. We might want to estimate the population standard deviation, the population median, or a population proportion (percentage). Using the COC work hours census data from fall 2015, we see that the population standard deviation is 12.969 hours per week. Remember, the two random sample standard deviations we have taken so far have both been off from that population standard deviation. Let us continue to collect random samples and calculate sample standard deviations. Again, we will take 251 random samples and calculate 251 random sample standard deviations. Each sample had a sample size of 50. If we put all of the sample standard deviations on the same graph, we can create a sampling distribution for sample standard deviations.



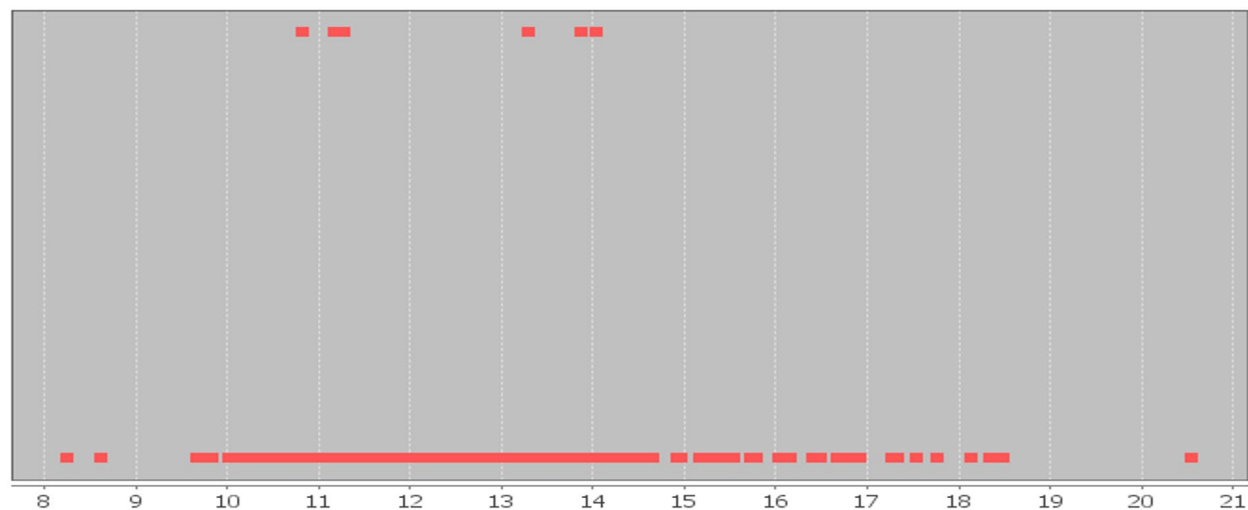
This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Population Parameters

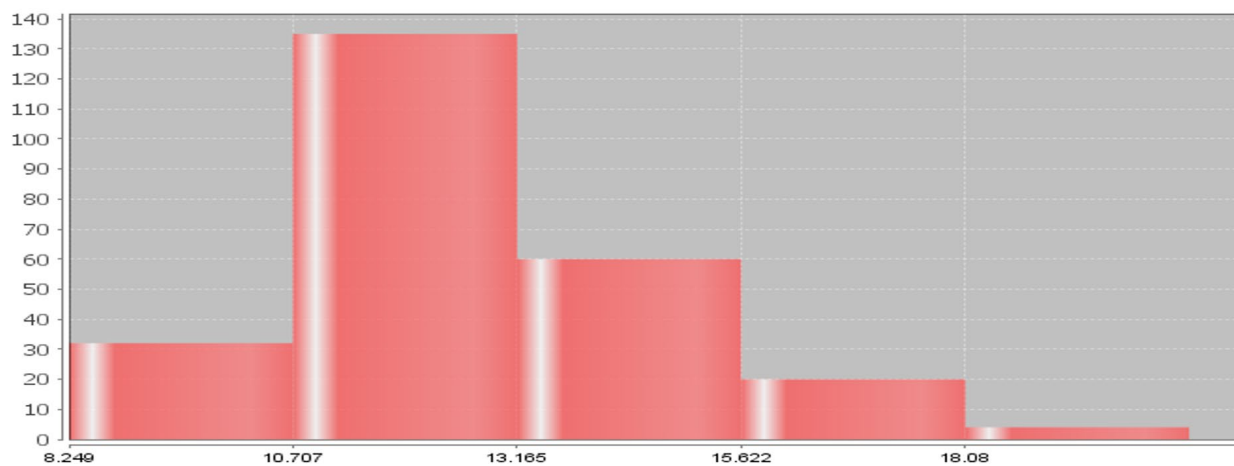
Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0

Histogram of Sampling Distribution for Sample Standard Deviations COC Work Hours Data



Histogram of Sampling Distribution for Sample Standard Deviations COC Work Hours Data



Descriptive Statistics

Variable	Center (Mean) of Sampling Distribution	Standard Error
Sampling Distribution for Sample Standard Deviations (COC stat students work hours)	12.636	1.998

Variable	Min	Max
Sampling Distribution for Sample Standard Deviations (COC stat students work hours)	8.249	20.538

Variable	Total Number of Random Samples
Sampling Distribution for Sample Standard Deviations (COC stat students work hours)	251

Notice that each dot in the sampling distribution represents the sample standard deviation of a random sample of size 50. We also created a histogram of the sampling distribution to judge shape. Notice a few things.

- We took 251 random samples and calculated 251 sample standard deviations. We see sampling variability in action. The population standard deviation is 12.969 hours per week but sample standard deviations ranged between 8.249 hours all the way to 20.538 hours. Random sample standard deviations are usually not the same as each other and usually very different from the population standard deviation (σ).
- Recall that the population was skewed right. The sampling distribution for these sample standard deviations also seems to have a skew. This can be a real problem. Remember the mean (center) and standard deviation (standard error) are not very accurate when data is not normal. For this reason, when estimating a population standard deviation, we like the population to be normal.
- Notice that the center (mean) of the sampling distribution is close to the population standard deviation of 12.969 hours per week. The mean average of all the sample standard deviations was 12.636 hours per week. The median average of all the sample standard deviations was 12.229. The median is a more accurate center since this sampling distribution was skewed, but remember standard error measures the distance to the mean of the sampling distribution, not the median.
- The standard error was 1.998. Remember, the standard error tells us how far typical sample statistics are from the center (mean) of the sampling distribution. Since the center of the sampling distribution is pretty close to the population value, the standard error tells us how far typical sample statistics are from the population parameter. In this case, it tells us that typical sample standard deviations are within 1.998 hours of the population standard deviation. Again, the accuracy of the center (mean) and the spread (standard error) are in question because the sampling distribution did not look normal.

Sampling Distributions for Sample Median Averages

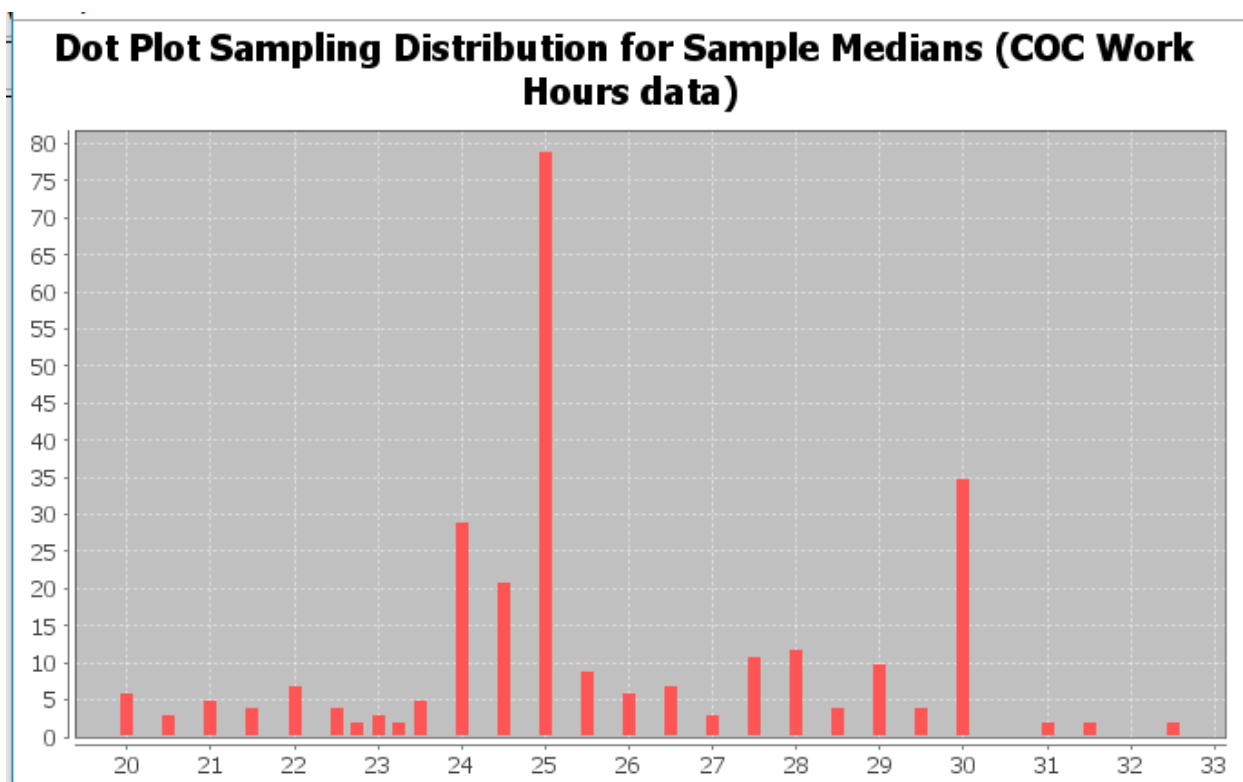
When data is skewed, we saw that the median average is usually more accurate than the mean, but how well do sample medians approximate population medians? Using the COC work hours census data from fall 2015, we see that the population median is 25 hours per week. Remember, the two random sample medians we have taken so far have both been off from that population median. Let us continue to collect random samples and calculate sample medians. Again, we will take 251 random samples and calculate 251 random sample medians. All of the samples had a sample size of 50. If we put all of the sample medians on the same graph, we can create a sampling distribution for sample medians.



Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0



Descriptive Statistics

Variable	Center (Mean) of Sampling Distribution	Standard Error
Sampling Distribution for Sample Medians (COC stat students work hours)	25.765	2.582

Variable	Min	Max
Sampling Distribution for Sample Medians (COC stat students work hours)	20.0	32.5

Variable	Total Number of Random Samples
Sampling Distribution for Sample Medians (COC stat students work hours)	251



Notice that each dot in the sampling distribution represents the sample median of a random sample. Notice a few things.

- We took 251 random samples and calculated 251 sample medians. We see sampling variability in action. The population median is 25 hours per week but sample medians ranged between 20 hours all the way to 32.5 hours. Random sample medians are usually not the same as each other and usually very different from the population median.
- Recall that the population was skewed right. The sampling distribution for these sample medians also seems to have a skew to the right. This again can be a real problem with the accuracy of the standard error.
- Again, we calculated the approximate center of the sampling distribution. This is the mean average of all of the sample medians. Notice that the center of the sampling distribution is 25.765 hours and is closer to the population median of 25 hours per week. Since this data was skewed to the right, the median of the sampling distribution will be a better measure of center. The median of the sampling distribution was 25 hours per week and in this case, was the same as the population median. Remember that the standard error measures the distance to the mean of the sampling distribution, not the median.
- We also calculated the standard error. Remember, the standard error tells us how far typical sample statistics are from the population parameter. In this case, it tells us that typical sample medians are within 2.582 hours of the population median. Again, the accuracy of the center (mean) and spread (standard error) are in question since the sampling distribution did not look normal.

Sampling Distributions for Sample Proportions (Sample Percentages)

Probably one of the most common population parameters that statisticians need to estimate is a population proportion or population percentage. There are important questions that need to be answered. What percentage of people in a country have health insurance? What percentage of people have diabetes?

To understand sampling variability for sample percentages we will again chose an example where we have census data and therefore know the population parameter. College of the Canyons (COC) has two campuses in the Santa Clarita Valley, the Valencia campus and the Canyon Country campus. We want to know what percentage of COC statistics students attend the Canyon Country campus. In 2015, we took a census of all of the statistics students at COC and found that the population percentage that attend the Canyon Country campus was 0.332 or 33.2%. If we take random samples of 40 students at a time from that population, will the sample proportions be 0.332? Let us find out.

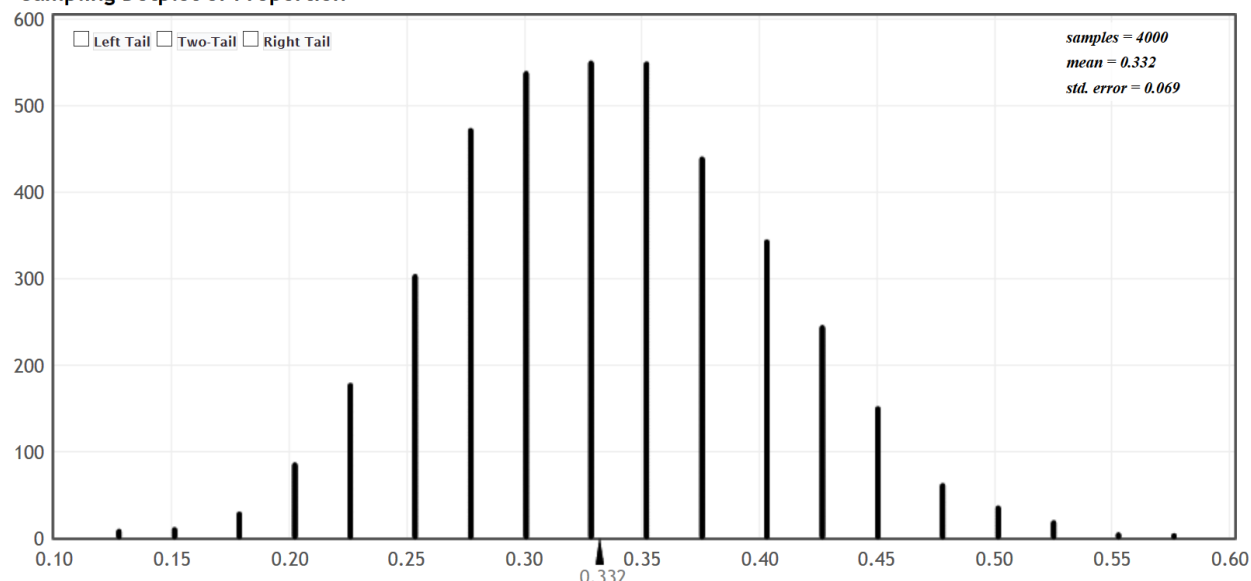
Here is a sampling distribution of thousands of random samples taken from the COC statistics student census. Remember the population proportion was 0.332.

Original Population

Proportion
0.332



Sampling Dotplot of Proportion



Notice that each dot in the sampling distribution represents the sample proportion of a random sample of 40 students. Notice a few things.

- We took 4000 random samples and calculated 4000 sample proportions. Again, we see sampling variability in action. The population proportion was 0.332 (33.2%) but sample proportions ranged between about 0.125 (12.5%) all the way to about 0.575 (57.5%). We see that there is a lot of sampling variability in sample proportions. Random sample proportions are usually not the same as each other and usually very different from the population proportion (π).
- Categorical data does not have a shape, but the sampling distribution for these sample proportions is normal.
- The center of the sampling distribution is calculated in the top right of the graph under “mean”. This is not the mean of a sample. It is the mean average of all the sample proportions. Notice that the center of the sampling distribution is 0.332 (33.2%) and is very close to the population proportion. In fact, the center of the sampling distribution is the same as the population proportion 0.332 (33.2%).
- In the top right of the graph you will again see “standard error”. Again, the standard error tells us how far typical sample statistics are from the center of the sampling distribution (population parameter). In this case it tells us that typical sample proportions are within 0.069 (6.9%) of the population proportion.

Key Notes about Sampling Distributions

1. Sampling Variability

Sampling distributions show us that random sample statistics are usually different from each other and different from the population parameter. Every time we take a random sample, we should expect to get different sample statistics and the statistics will be off from the population parameter.

2. Shape of Sampling Distributions

The shape of sampling distribution is very important. Remember the center (mean) and spread (standard error) of the sampling distribution are only accurate if the sampling distribution is normal. We saw that if the population is skewed, the sampling distribution may or may not be normal. This important topic needs further exploration.

3. Population Parameter \approx Center of the Sampling Distributions

While one sample statistic can be far off from the population parameter, the center of a sampling distribution is usually very close to the population parameter. Let us suppose you are in a situation where you cannot collect an unbiased census. If you are able to collect multiple random samples, you can start to create a sampling distribution.



This chapter is from *Introduction to Statistics for Community College Students*,
 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
 under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Then look for the center of the distribution and you will usually have a good approximation of the population parameter. If you are using the mean of the sampling distribution as the center, we will want the sampling distribution to be normal.

Political election polls are usually dramatically off from what will happen on voting day. Yet as we get closer and closer to voting day, statisticians and data scientists seem to have a better idea of how the voting will go? If we base our population percentage of voting on one sample (one poll), we may be very far off. By the time of the vote, we have taken many polls, many samples. If we put all the sample percentages on the same graph, we have created a sampling distribution for sample proportions. Go to the center of the graph and you will have a much better idea of the population proportion, the population percentage of who will vote in what direction.

4. Standard Error and Margin of Error

Standard error is the standard deviation of the sampling distribution and tells us how far typical statistics could be from the population parameter. The accuracy of the standard error is highly reliant on the sampling distribution being normal.

Remember standard error and margin of error are not the same thing. Standard error measures typical statistics. Many sample statistics may not be typical. The margin of error considers sample statistics that are not just typical. Usually the margin of error is about twice as large as the standard error.

Optional Sampling Distribution Class Activity 1

Exploring Sampling Variability for Mean Averages with a Sampling Distribution

The goal of this activity is to explore how well random samples approximate population values. Normally we do not know population values and we must use a sample value to approximate the population value. This is called a “point estimate”. For this activity, we will look at some population data from International Coffee Organization (ICO). We will be using the “Columbian Mild” price data in U.S. cents per pound. The population mean average price was 136.43 cents per pound. Again, in real data analysis, we often do not know the population value, but for this activity, it is useful for comparison purposes.

Open the “Sampling Distribution Data 1” in Excel. A total of 120 random samples have been taken from the Columbian Mild data. All the data sets have 30 coffee prices. Each person in the class will be finding the mean of a few of these data sets. Once you find your sample means, you will put a magnet up or draw a dot on the board to represent the sample mean you found. When everyone’s magnets or dots are up on the board, we will have generated a “sampling distribution”.

Answer the following questions:

1. The population mean was 136.43 cents. How many cents was the sample mean you calculated from the population mean of 136.43 cents? (If you calculated more than one sample mean, answer the question for all the sample means you calculated.) This is called the “Margin of Error”.
2. Look at the dots or magnets on the board. Did all the sample means come out to be the same as the population mean of 136.43cents? Why do you think this happened? Aren’t random samples supposed to be good approximations of the population? What does this tell you about sampling variability?
3. Normally, we may have only one random sample. If all you knew was one of the random samples on the board, how difficult would it be to determine that the population mean is really 136.43 cents? What does this tell us about the difficulty in determining population values from one random sample?
4. Estimate the shape and center of the sampling distribution on the board. Is the center of the graph close to the population mean of 136.43? Would the center of the sampling distribution be a better approximation of the population mean than a single sample mean?



5. The standard deviation of a sampling distribution is often called the “standard error” and is an important part of inferential statistics. Estimate how far typical dots are from the center of the sampling distribution. This is the standard deviation of the sampling distribution, which is called “Standard Error”.

Optional Sampling Distribution Class Activity 2

Exploring Sampling Variability for Percentages with a Sampling Distribution

The goal of this activity is to explore how well random sample percentages approximate population percentages. Normally we do not know population percentage and we must use a sample percentage to approximate the population percentage. This is called a “point estimate”. For this activity, we will be flipping coins 30 times and count the number of tails. Then calculate the sample percentage of tails. Each person will do three sets of 30 and therefore get three sample percentages. Again, in real data analysis, we often do not know the population value, but for this activity, it is useful for comparison purposes. Our goal is to see how well random sample percentages approximate population percentages.

Each person in the class will be finding three sample percentages. Once you find each sample percent, you will put a magnet up or draw a dot on the board to represent the sample percent you found. When everyone’s magnets or dots are up on the board, we will have generated a “sampling distribution” of sample percentages.

Answer the following questions:

1. In a perfect world and a fair coin, what should the population percentage for getting tails be? So in a sample of 30 how many times do we expect to get tails? In sampling, we often do not get what we expect. How far were the sample percentages you calculated from the population percentage?
 2. Look at the dots or magnets on the board. Did all the sample percentages come out to be the same as the population percentage? Why do you think this happened? Aren’t random samples supposed to be good approximations of the population? What does this tell you about sampling variability?
 3. Normally, we may have only one random sample. If all you knew was one of the sample percentage on the board, and you never knew the expected population value, how difficult would it be to determine what the population percentage really is? What does this tell us about the difficulty in determining population values from one random sample?
 4. Estimate the shape and center of the sampling distribution on the board. Is the center of the graph close to the population percentage of 0.5? Would the center of the sampling distribution be a better approximation of the population percentage than a single sample percentage?
 5. The standard deviation of a sampling distribution is often called the “standard error” and is an important part of inferential statistics. Estimate how far typical dots are from the center of the sampling distribution. This is the standard deviation of the sampling distribution, which is called “Standard Error”.
-

Problem Set Section 2B

Directions: Answer the following questions about sampling distributions.

1. Describe the process of making a sampling distribution.
2. What can sampling distributions tell us about sampling variability?
3. What is a point estimate? Discuss how point estimates create confusion for people reading articles and scientific reports.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-BY” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

4. Discuss the shape of sampling distributions. When the population is skewed, is the sampling distribution always normal? Why is it important for a sampling distribution to be normal? In the examples in this section, which statistics had a normal sampling distribution? Which statistics had a skewed sampling distribution?

5. Explain how the standard error is calculated. What does the standard error tell us about sample statistics and the population parameter? Why is the standard error only accurate when the sampling distribution is normal?

6. What is the difference between standard error and margin of error? Is the standard error smaller or larger than the margin of error?

(#7-16) For the following problems, copy the indicated census data set from the Math 140 Survey Data at www.matt-teachout.org. We will be assuming this is an unbiased census and therefore know the population mean. Open StatKey at www.lock5stat.com. Under the “sampling distributions” menu, click on “mean”. You should see “sampling distribution for the mean”. Under “edit data” paste in the indicated data set. Under “choose samples of size n”, put in the indicated sample size. Create a sampling distribution and then answer the following questions.

7. Use StatKey to create a sampling distribution with sample size 10 from the Age in Years census data (Math 140 Survey Data).

- What was the shape and mean average of the population?
- Were all the sample means the same as the population mean?
- Were all the sample means the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- What is the standard error? Write a sentence explaining the meaning of the standard error.

8. Use StatKey to create a sampling distribution with sample size 100 from the Age in Years census data (Math 140 Survey Data)

- What was the shape and mean average of the population?
- Were all the sample means the same as the population mean?
- Were all the sample means the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- What is the standard error? Write a sentence explaining the meaning of the standard error.
- How does the standard error for sample size 10 compare to the standard error for sample size 100?
- How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

9. Use StatKey to create a sampling distribution with sample size 10 from the sleep hours per night census data (Math 140 Survey Data)

- What was the shape and mean average of the population?
- Were all the sample means the same as the population mean?
- Were all the sample means the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- What is the standard error? Write a sentence explaining the meaning of the standard error.

10. Use StatKey to create a sampling distribution with sample size 25 from the sleep hours per night census data (Math 140 Survey Data)

- What was the shape and mean average of the population?
- Were all the sample means the same as the population mean?
- Were all the sample means the same as each other?
- How many random samples did you take when you created the sampling distribution?



- e) What is the shape of the sampling distribution?
- f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- g) What is the standard error? Write a sentence explaining the meaning of the standard error.
- h) How does the standard error for sample size 10 compare to the standard error for sample size 25?
- i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 25?

11. Use StatKey to create a sampling distribution with sample size 10 from the cell phone bill (in dollars per month) census data (Math 140 Survey Data).

- a) What was the shape and mean average of the population?
- b) Were all the sample means the same as the population mean?
- c) Were all the sample means the same as each other?
- d) How many random samples did you take when you created the sampling distribution?
- e) What is the shape of the sampling distribution?
- f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- g) What is the standard error? Write a sentence explaining the meaning of the standard error.

12. Use StatKey to create a sampling distribution with sample size 100 from the cell phone bill (in dollars per month) census data (Math 140 Survey Data).

- a) What was the shape and mean average of the population?
- b) Were all the sample means the same as the population mean?
- c) Were all the sample means the same as each other?
- d) How many random samples did you take when you created the sampling distribution?
- e) What is the shape of the sampling distribution?
- f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- g) What is the standard error? Write a sentence explaining the meaning of the standard error.
- h) How does the standard error for sample size 10 compare to the standard error for sample size 100?
- i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

13. Use StatKey to create a sampling distribution with sample size 10 from the travel time to get to school in minutes (Math 140 Survey Data).

- a) What was the shape and mean average of the population?
- b) Were all the sample means the same as the population mean?
- c) Were all the sample means the same as each other?
- d) How many random samples did you take when you created the sampling distribution?
- e) What is the shape of the sampling distribution?
- f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- g) What is the standard error? Write a sentence explaining the meaning of the standard error.

14. Use StatKey to create a sampling distribution with sample size 40 from the travel time to get to school in minutes (Math 140 Survey Data).

- a) What was the shape and mean average of the population?
- b) Were all the sample means the same as the population mean?
- c) Were all the sample means the same as each other?
- d) How many random samples did you take when you created the sampling distribution?
- e) What is the shape of the sampling distribution?
- f) What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- g) What is the standard error? Write a sentence explaining the meaning of the standard error.
- h) How does the standard error for sample size 10 compare to the standard error for sample size 40?
- i) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 40?



15. Use StatKey to create a sampling distribution with sample size 10 from the work hours per week for COC college students (Math 140 Survey Data).

- What was the shape and mean average of the population?
- Were all the sample means the same as the population mean?
- Were all the sample means the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- What is the standard error? Write a sentence explaining the meaning of the standard error.

16. Use StatKey to create a sampling distribution with sample size 40 from the work hours per week for COC college students (Math 140 Survey Data).

- What was the shape and mean average of the population?
- Were all the sample means the same as the population mean?
- Were all the sample means the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of the sampling distribution? Is it relatively close to the population mean?
- What is the standard error? Write a sentence explaining the meaning of the standard error.
- How does the standard error for sample size 10 compare to the standard error for sample size 40?
- How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 40?

(#17-26) The following population proportions come from the Math 140 Survey Data at www.matt-teachout.org. We will be assuming this is an unbiased census and therefore know the population proportion (%). Open StatKey at www.lock5stat.com. Under the "sampling distributions" menu, click on "proportion". You should see "sampling distribution for a proportion". Under "edit proportion", enter the given population proportion. Create a sampling distribution and then answer the following questions.

17. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students with brown hair is 0.537. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

- Were all the sample proportions the same as the population proportion?
- Were all the sample proportions the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- What is the standard error? Write a sentence explaining the meaning of the standard error.

18. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students with brown hair is 0.537. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

- Were all the sample proportions the same as the population proportion?
- Were all the sample proportions the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- What is the standard error? Write a sentence explaining the meaning of the standard error.
- How does the standard error for sample size 10 compare to the standard error for sample size 100?
- How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?



19. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students that smoke cigarettes is 0.091. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

- Were all the sample proportions the same as the population proportion?
- Were all the sample proportions the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- What is the standard error? Write a sentence explaining the meaning of the standard error.

20. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students that smoke cigarettes is 0.091. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

- Were all the sample proportions the same as the population proportion?
- Were all the sample proportions the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- What is the standard error? Write a sentence explaining the meaning of the standard error.
- How does the standard error for sample size 10 compare to the standard error for sample size 100?
- How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

21. Approximately 60% of college students in the U.S. were able to finish their bachelor's degree in six years. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

- Were all the sample proportions the same as the population proportion?
- Were all the sample proportions the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- What is the standard error? Write a sentence explaining the meaning of the standard error.

22. Approximately 60% of college students in the U.S. were able to finish their bachelor's degree in six years. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

- Were all the sample proportions the same as the population proportion?
- Were all the sample proportions the same as each other?
- How many random samples did you take when you created the sampling distribution?
- What is the shape of the sampling distribution?
- What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- What is the standard error? Write a sentence explaining the meaning of the standard error.
- How does the standard error for sample size 10 compare to the standard error for sample size 100?
- How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

23. Approximately 9.4% of all adults in the U.S. have diabetes. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

- Were all the sample proportions the same as the population proportion?
- Were all the sample proportions the same as each other?
- How many random samples did you take when you created the sampling distribution?



- d) What is the shape of the sampling distribution?
- e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- f) What is the standard error? Write a sentence explaining the meaning of the standard error.

24. Approximately 9.4% of all adults in the U.S. have diabetes. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

- a) Were all the sample proportions the same as the population proportion?
- b) Were all the sample proportions the same as each other?
- c) How many random samples did you take when you created the sampling distribution?
- d) What is the shape of the sampling distribution?
- e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- f) What is the standard error? Write a sentence explaining the meaning of the standard error.
- g) How does the standard error for sample size 10 compare to the standard error for sample size 100?
- h) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

25. Approximately 90% of all lung cancer cases are caused by cigarette smoking. Use this population proportion to create a sampling distribution with sample size 10 with StatKey.

- a) Were all the sample proportions the same as the population proportion?
- b) Were all the sample proportions the same as each other?
- c) How many random samples did you take when you created the sampling distribution?
- d) What is the shape of the sampling distribution?
- e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- f) What is the standard error? Write a sentence explaining the meaning of the standard error.

26. Approximately 90% of all lung cancer cases are caused by cigarette smoking. Use this population proportion to create a sampling distribution with sample size 100 with StatKey.

- a) Were all the sample proportions the same as the population proportion?
- b) Were all the sample proportions the same as each other?
- c) How many random samples did you take when you created the sampling distribution?
- d) What is the shape of the sampling distribution?
- e) What is the center (mean) of all the sample proportions in the sampling distribution? Is it relatively close to the population proportion (π)?
- f) What is the standard error? Write a sentence explaining the meaning of the standard error.
- g) How does the standard error for sample size 10 compare to the standard error for sample size 100?
- h) How does the shape of the sampling distribution for sample size 10 compare to the shape of the sampling distribution for sample size 100?

Section 2C – The Central Limit Theorem

In the last section, we saw that when estimating population parameters from samples, it is very important for a sampling distribution to be normal. The accuracy of the center of the sampling distribution (population estimate) and the spread of the sampling distribution (standard error) are tied to the sampling distribution being normal. We also saw that if the population was skewed, the sampling distribution may or may not look normal. In this section, we will discuss further the shape of sampling distributions and determine what conditions need to be met in order to get a normal sampling distribution.

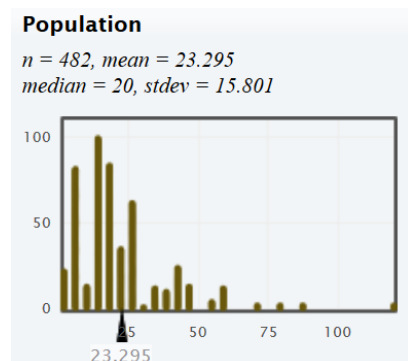
Sample Means



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Let us start by looking at sample means. Let us look at the census of College of the Canyons (COC) statistics students taken in the fall 2015 semester. The variable we will look at is how many minutes it takes to commute to COC.

Census Data (Commute Time in Minutes for COC Stat Students Fall 2015)



We see that the population is skewed with a population mean average commute time of 23.295 minutes. We will assume that the census is unbiased and that the population mean is really 23.295 minutes.

Key Question: If the population is skewed, what conditions need to be met in order for the sampling distribution to look normal?

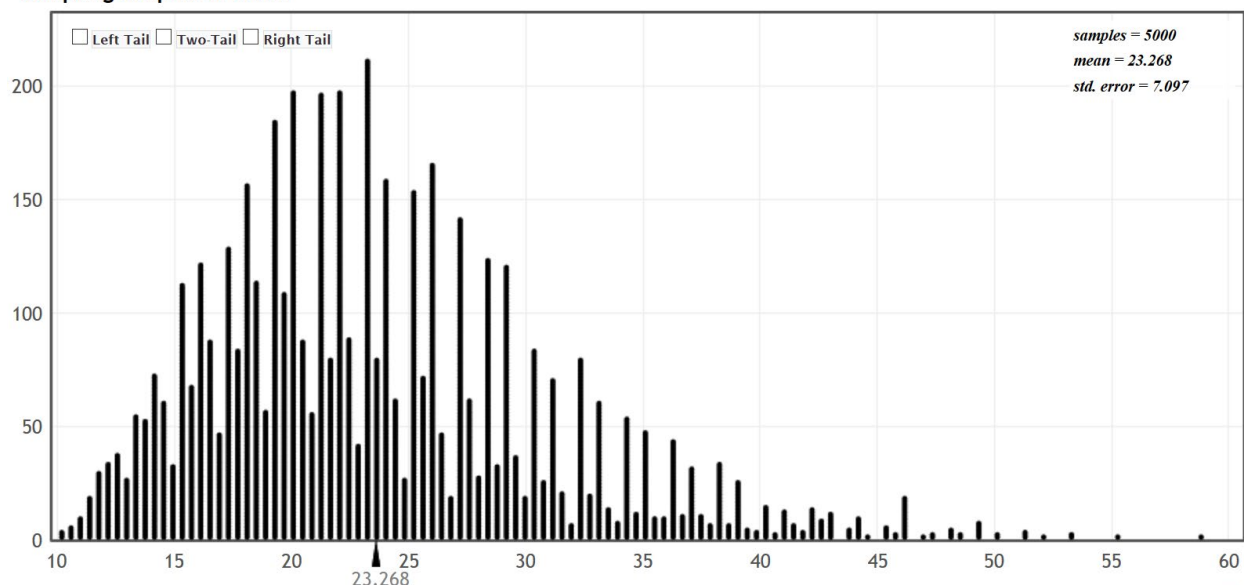
Mean Example 1: Sample Size of Seven from a Skewed Population

Let us take many random samples from the census of COC stat students commute times, calculated the sample mean from each sample and then put the sample means on the same graph. This is called a sampling distribution for sample means. For this example, we used small samples with a sample size of seven. We used the sampling distribution function on StatKey to create 5000 random samples with each sample have seven commute times. We calculated 5000 sample means and put them on the same graph. Notice the sampling distribution still looks skewed. In addition, notice that the center (mean) of the sampling distribution was 23.268 minutes and the standard error is 7.097 minutes. We would not trust the accuracy of the standard error or the mean of the sampling distribution because the sampling distribution was not normal.

- Shape of Sampling Distribution: Skewed Right
- Center (mean) of the sampling distribution ≈ 23.268 minutes
- Standard error ≈ 7.097 minutes.



Sampling Dotplot of Mean



Mean Example 2: Sample Size of Twenty-Five from a Skewed Population

Let us create another sampling distribution from the census of COC stat student commute times. This time we will increase the sample size to twenty-five. Each sample will have twenty-five commute times. We used the sampling distribution function on StatKey to create 5000 random samples with each sample having a sample size of twenty-five. Notice the sampling distribution now looks nearly normal. The center (mean) of the sampling distribution was 23.336 minutes and the standard error is 3.108 minutes. We can trust the accuracy of the standard error and the mean of the sampling distribution because the sampling distribution was nearly normal.

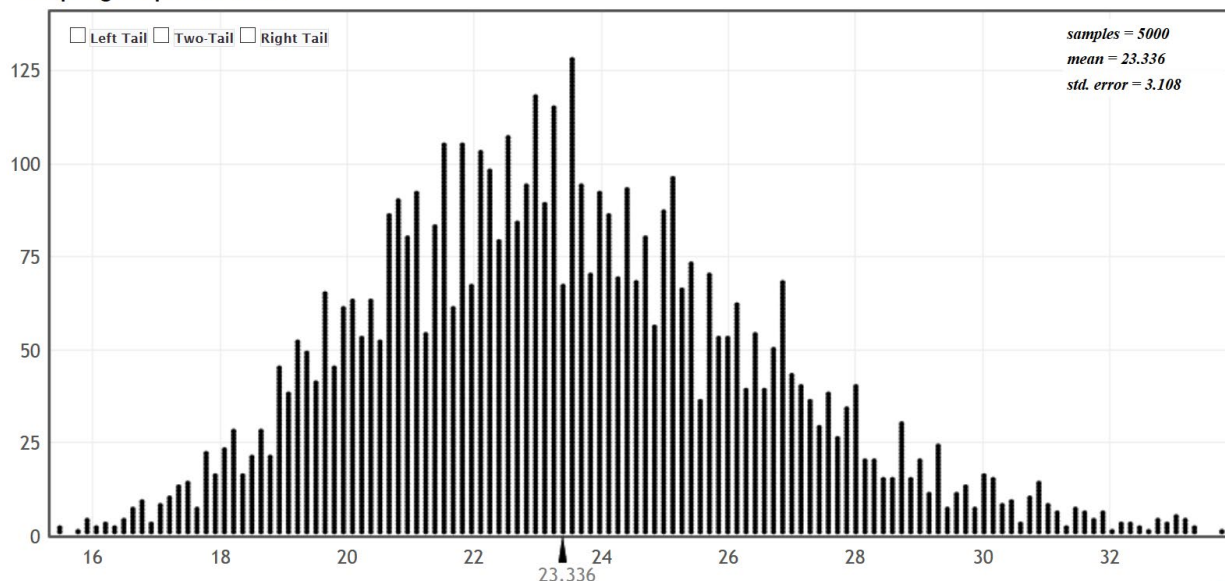
- Shape of Sampling Distribution: Nearly Normal
- Center (mean) of the sampling distribution ≈ 23.336 minutes
- Standard error ≈ 3.108 minutes.

Notice also that the standard error for this sample size ($n=25$) is smaller than the standard error for the very small sample size ($n=7$). This is a very important principle, more random data results in less error. The larger the sample size, the smaller the standard error, and the more normal the sampling distribution looks.

More Random Data \rightarrow Less Error \rightarrow Sampling Distribution becomes more normal



Sampling Dotplot of Mean



Mean Example 3: Sample Size of Two Hundred from a Skewed Population

Let us create one more sampling distribution from the COC stat students commute times data. This time we will increase the sample sizes to two hundred. Notice the sampling distribution now looks very normal. The center (mean) of the sampling distribution was 23.290 minutes and the standard error has dropped to 0.863 minutes.

- Shape of Sampling Distribution: Normal
- Center (mean) of the sampling distribution ≈ 23.290 minutes
- Standard error ≈ 0.863 minutes.

Notice also that the standard error for this sample size ($n=200$) is smaller than the standard error for the sample size ($n=25$). Remember, the larger the sample size, the smaller the standard error, and the more normal the sampling distribution looks.

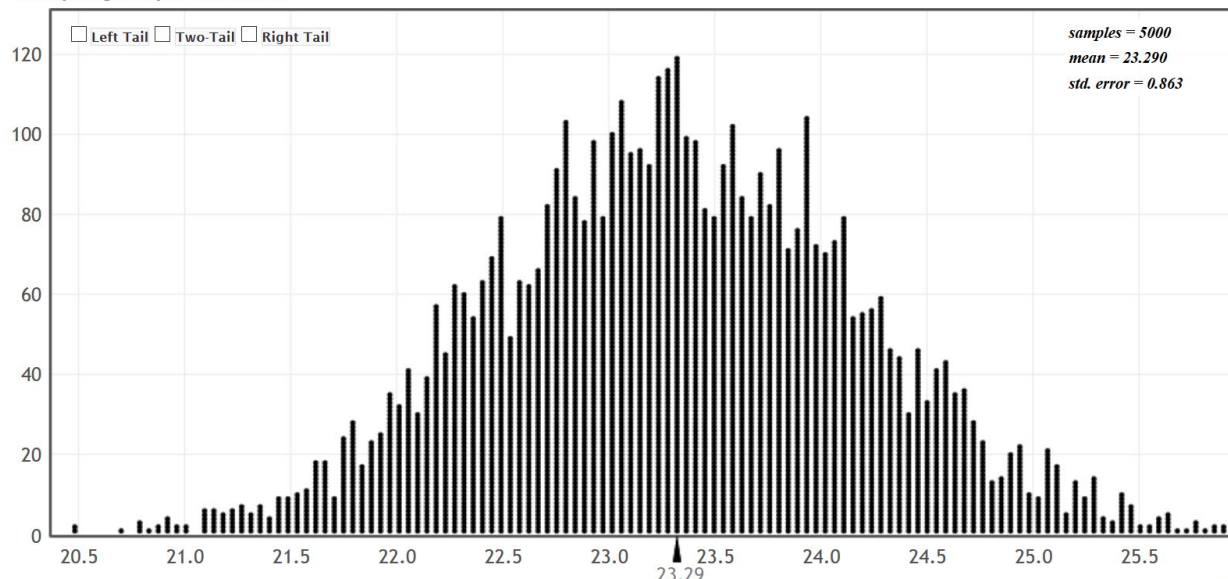
More Random Data \rightarrow Less Error \rightarrow Sampling Distribution becomes more normal



StatKey Sampling Distribution for a Mean

Custom Dataset ▾ Show Data Table Edit Data Choose samples of size $n =$ 200 Upload File Change Column(s)
 Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Sampling Dotplot of Mean



Summary: If a population is skewed, it seems we need a larger sample size, for the sampling distribution to look normal. As the sample size increases, the standard error decreases, and the sampling distribution looks more normal. This is the idea behind the “Central Limit Theorem”. A common rule when dealing with means is that if the population is skewed the sample size should be at least 30 for the sampling distribution for sample means to look normal.

Central Limit Theorem: If the sample size is sufficiently large, the sampling distribution for sample means will have a normal shape even if the population is skewed.

Key Question: What would happen if the population were already normal?

Mean Example 4: Sampling Distribution from a normal population.

Let us now look at an example of a census with a normal shape. In the fall 2015 semester, we took a census of all of the statistics students at COC and asked them their heights in inches. We will assume this was an unbiased census. This population looked very normal with a population mean average height of 66.511 inches and a population standard deviation of 4.787 inches. For this example, we will focus on the mean.

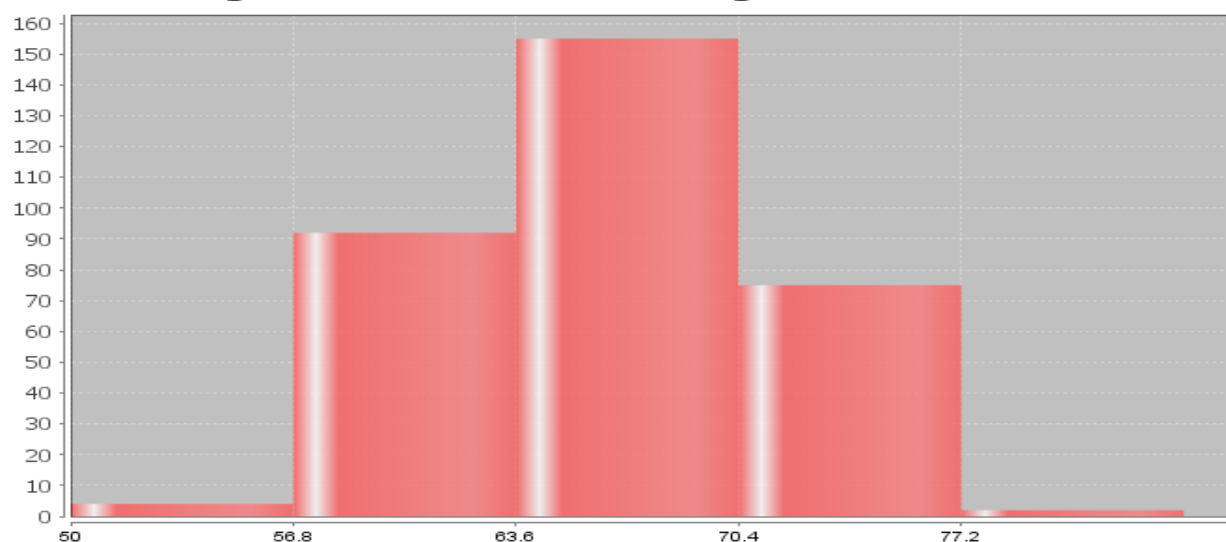
Population Parameters

Variable	Population Mean	Population Standard Deviation
COC Stat Students Population height (in INCHES)	66.511	4.787

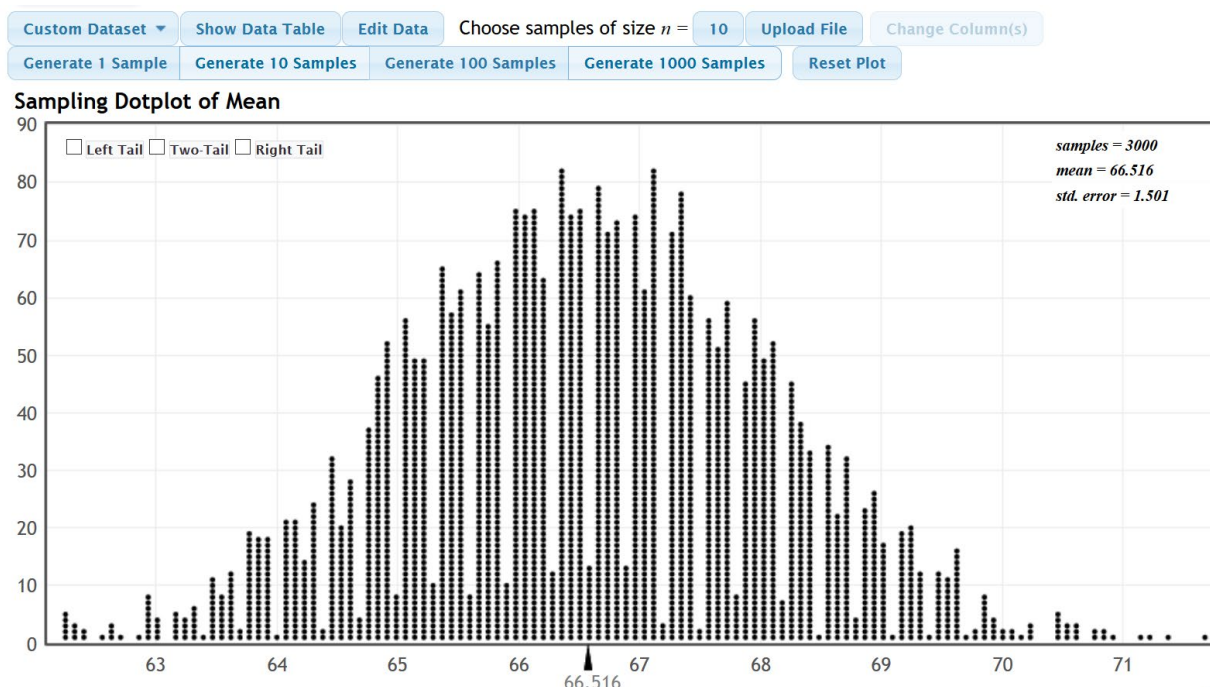


This chapter is from *Introduction to Statistics for Community College Students*,
 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
 under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Histogram COC Stat Student Height Fall 2015 Census



Now let us see what happens if we take thousands of samples from this population. We will start with small sample sizes of 10 stat students at a time.



We took 3000 random samples each of size ten and calculated 3000 sample means to create this sampling distribution. Notice a few key things.

- The sample means are different. We see sampling variability in action. The population mean was 66.511 inches but the sample means could be anywhere from about 62 inches to 72 inches. Sample statistics are different and usually very different than the population parameter.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

- Even though we have a very small sample size of ten, the sampling distribution still looks normal. This means that the center (mean) of the sampling distribution and the standard error are relatively accurate even for a sample size of ten.
- The center of the sampling distribution (66.516 inches) is very close to the population mean of (66.511 inches)
- We have calculated the standard error of 1.501. For a sample size of 10, typical sample means are within 1.501 inches of the population mean. The margin of error is probably closer to 3 inches (2 x standard error).

Sample Mean Summary

Let us summarize our findings about sample means from random samples.

1. If the population is skewed, we will need a sample size of at least 30 or higher in order to insure that our sampling distribution for sample means will be nearly normal.
2. If the population is already normal, then the sampling distribution for sample means will be normal for any sample size.

Important note about sample size:

Even though the minimum requirement for sample means is a sample size of 30 or above, this does not mean we are happy with a data set of only 30. Remember less data results in more error. For random data, the bigger the sample size the better. Thirty is just the bare minimum requirement to insure that the sampling distribution for sample means will look nearly normal.

Standard Deviation Example 1: Standard Deviation and Variance

Remember that the sample variance is the square of the standard deviation. Statisticians often opt to estimate variability in sample variances instead of standard deviation. Later, we can take the square root of the variance estimates to get the standard deviation.

If the population was skewed, what is the shape of sampling distributions for sample standard deviations? Are there any sample size requirements for estimating sample standard deviations? What if the population was already normal?

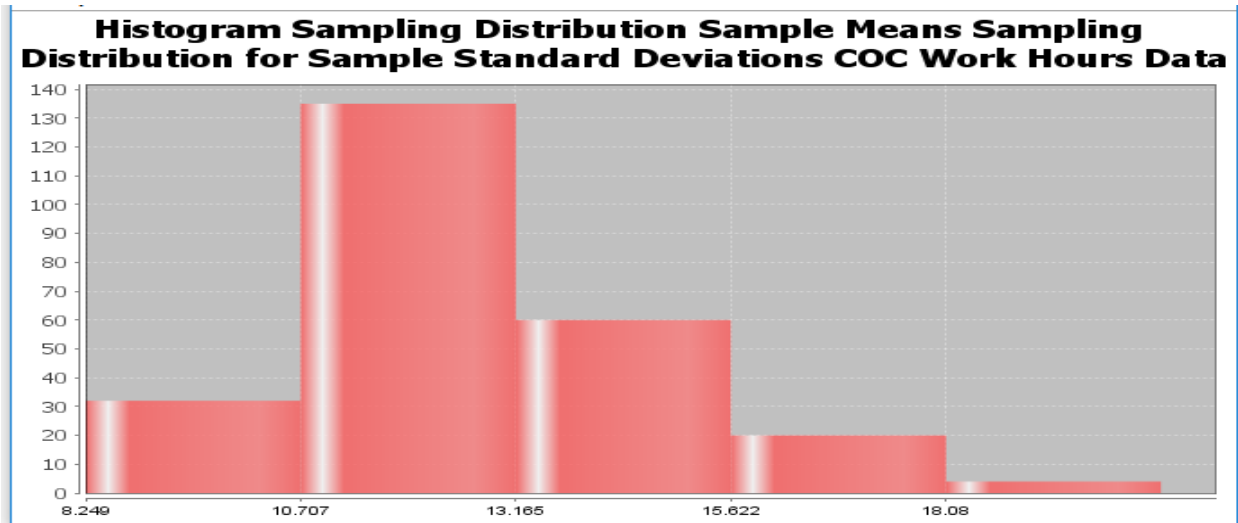
In the last section, we looked at the COC work hours census data from fall 2015. We see that the population standard deviation is 12.969 hours per week. We created a sampling distribution of 251 random samples and calculate 251 random sample standard deviations. Each sample had a sample size of 50. If we put all of the sample standard deviations on the same graph, we can create a sampling distribution for sample standard deviations.

Population Parameters

Variable	Mean	Standard Deviation
Work Hours per Week COC Stat Students	27.283	12.969

Variable	Median
Work Hours per Week COC Stat Students	25.0





Notice that while a sample size of 50 would be large enough to ensure a sampling distribution of sample means to be normal; it does not insure a sampling distribution of sample standard deviations to be normal. If the population is skewed, the sampling distribution for sample standard deviations will tend to be skewed.

Sample Standard Deviation and Sample Variance Summary

Let us summarize our findings about sample standard deviations and sample variance from random samples.

1. A sampling distributions of sample variance is usually skewed right. Later we will see that if the population is normal, the sampling distribution for sample variance will follow a skewed right Chi-Squared distribution. Requirements for traditional techniques for estimating population variance or population standard deviations usually require the population to be normal no matter what the sample size is. If the population were not normal, then we would have to resort to different technique like bootstrapping.

Sample Proportion Example 1:

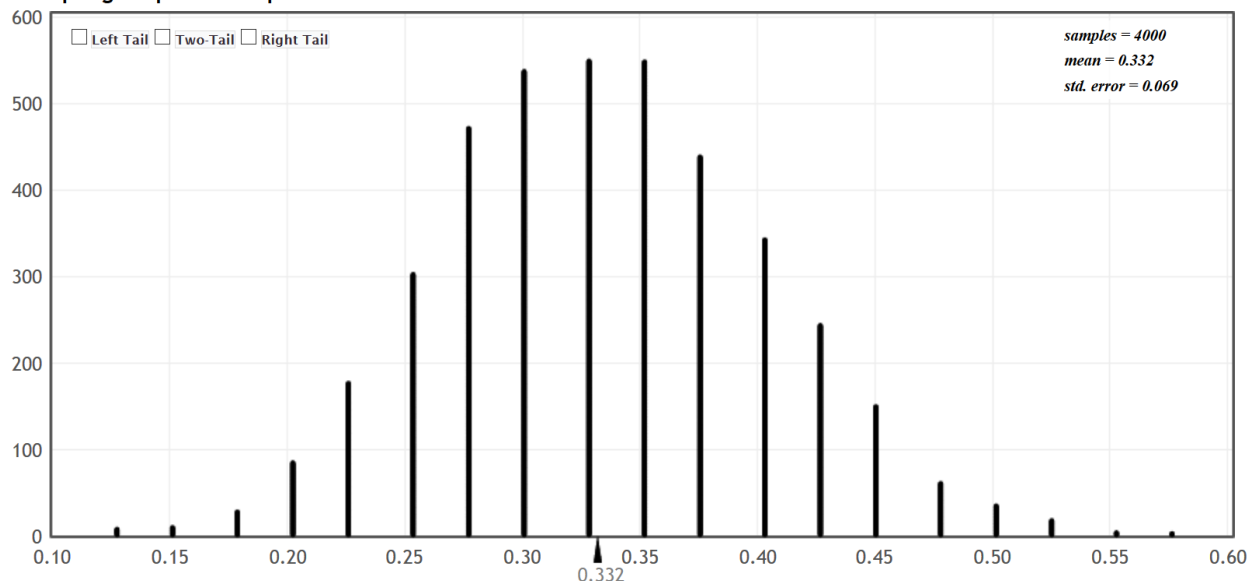
In the last section, we looked at the fall 2015 census of COC stat students and found that the population percentage that attend the Canyon Country campus was 0.332 or 33.2%. Here is a sampling distribution of thousands of random samples taken from the COC statistics student census. Remember the population proportion was 0.332.

Original Population

Proportion
0.332



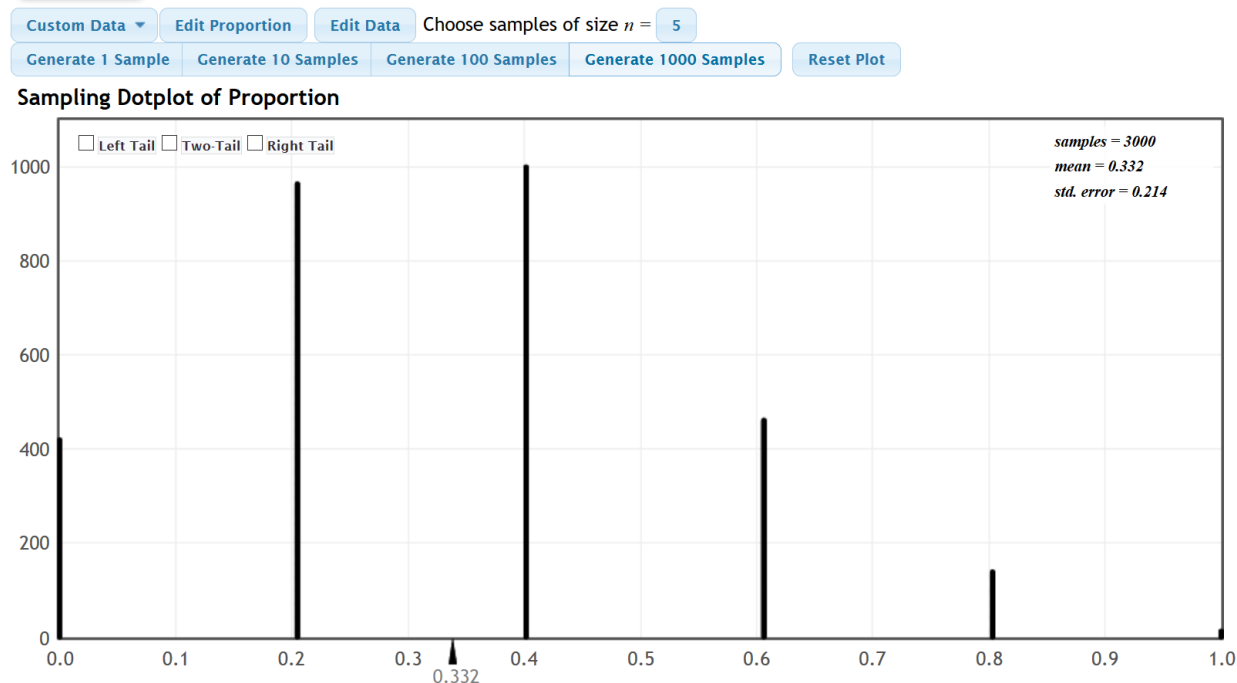
Sampling Dotplot of Proportion



Notice that for a sample size of 40, the sampling distribution looks normal. In addition, the center of the sampling distribution was very close to the population proportion of 0.332.

What if we decrease the sample size?

Let us look at a sampling distribution for sample proportions from the same population, but now we will decrease the sample size to five.



Notice at a sample size of only five, the sampling distribution looks skewed right. Notice that the center (mean) of all the sample proportions was still very close to 0.332, but we will not have much confidence in the standard error from a sampling distribution that is not normal.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

So what sample size insures a normal sampling distribution for sample proportions?

The “At Least Ten” rule

It turns out that for random categorical data, the random sample should have at least ten successes and at least ten failures. We should have at least 10 statistics students from the Canyon Country campus and at least 10 that are not from the Canyon Country campus to insure that the sampling distribution will look normal.

Notice if we only had a random sample of five stat students, it is impossible to get at least ten from Canyon Country and at least ten not from Canyon Country.

There is no minimum sample size requirement for categorical data because the population proportion will be different in each situation.

Why did the sampling distribution for samples of size 40 work?

If we know the population proportion (π), here is a common formula for estimating the number of success and failures in random categorical sample data:

Expected number of success for sample size (n): $n(\pi)$

Expected number of failures for sample size (n): $n(1 - \pi)$

For a sample size of 40, will we be likely to get ten successes and ten failures? If the population proportion for Canyon Country is 0.332, we are likely to get about 13 students from Canyon Country and 27 students not from Canyon Country.

$$n(\pi) = 40(0.332) = 13.28$$

$$n(1 - \pi) = 40(1 - 0.332) = 40(0.668) = 26.72$$

Important Note: Remember we rarely have an unbiased census, so we may have no idea what the population proportion is. All we have is random sample data. In that case, you will want your random categorical sample data to have at least ten success and at least ten failures. That does not mean twenty!

Summary of Sampling Distributions for sample proportions (sample %)

- Categorical data does not have a shape. Yet if we compute thousands of sample proportions and put them on the same graph, the sampling distribution will have a shape.
- To insure the sampling distribution for sample proportions will be normal we want to have at least ten successes and at least ten failures in our random categorical sample data.

Key Question#1: Why is it so important for a sampling distribution to be normal?

We will discuss this in greater detail in later sections, but here are two of the main reasons.

- Remember standard error is the standard deviation of the sampling distribution and measures the typical distance from the mean (center) of the sampling distribution. Neither the standard error nor the center (mean) of the sampling distribution are very accurate unless the sampling distribution is normal.
- Before computers were invented, statisticians relied on formulas to understand sampling variability, calculate standard error and estimate population parameters. Many of these formulas are based on normal curves and are not accurate if the sampling distribution is not normal. This is why conditions or assumptions for sample means and sample proportions are often tied to making sure the sampling distribution is normal when estimating population parameters.



Key Question#2: Is there a way to estimate a population parameter and understand sampling variability when the sampling distribution is not normal?

- Yes. Computer technology may be used to understand sampling variability in the case when our sampling distribution is not likely to be normal. Techniques like bootstrapping and randomized simulation were invented to be able to understand sampling variability, calculate standard error, and estimate or check population parameters when the sampling distribution is not normal. We will discuss these techniques in later chapters.
-

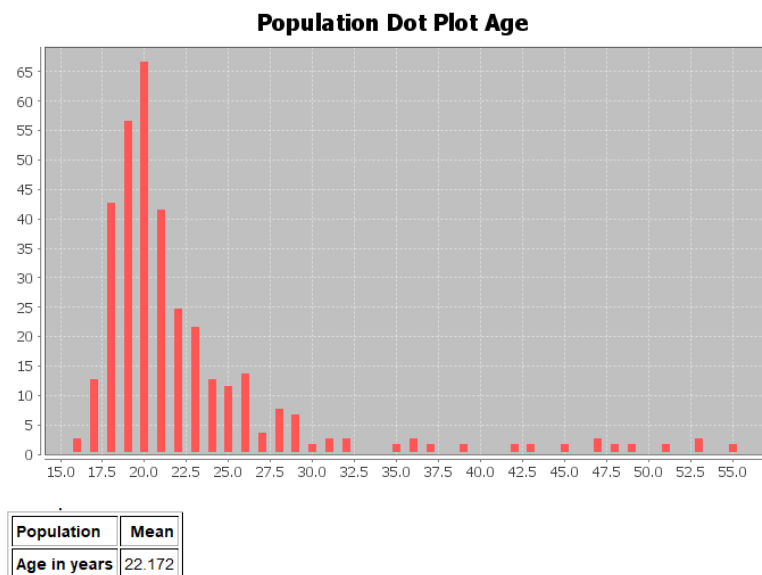
Problem Set Section 2C

Directions: Answer the following questions about sample size requirements and the shape of sampling distributions.

1. Why is it important for a sampling distribution for sample means or sample proportions to be normal?
2. What conditions should be met to insure that a sampling distribution of sample proportions is normal?
3. State the Central Limit Theorem and explain the ideas behind it.
4. Suppose the population is not normal. If we increase the sample size, what will happen to the standard error and the shape of the sampling distribution of sample means?
5. Suppose the population is not normal. If we decrease the sample size, what will happen to the standard error and the shape of the sampling distribution of sample means?
6. Suppose the population is not normal. What conditions should be met in order to insure that a sampling distribution of sample means is normal?
7. If the population is normal, will the sampling distribution for sample means look normal for very small sample sizes?
8. Median averages, variance and standard deviation can have very irregular looking sampling distributions. This can make traditional formula calculations difficult. Is there a way to study sampling variability and estimate population parameters when a sampling distribution is not normal or when traditional formulas are not accurate?

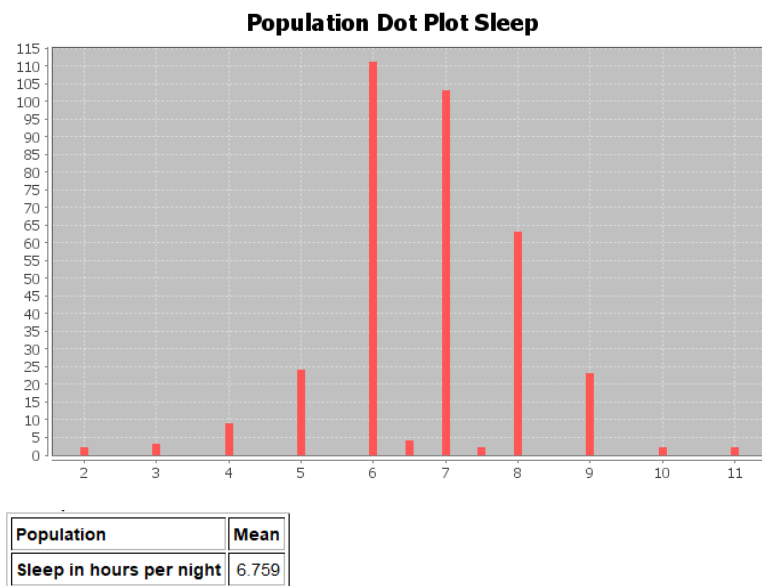


9. The following graph and population mean were created with Statcato from the “age in years” census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



- What was the shape and mean average of the population?
- If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

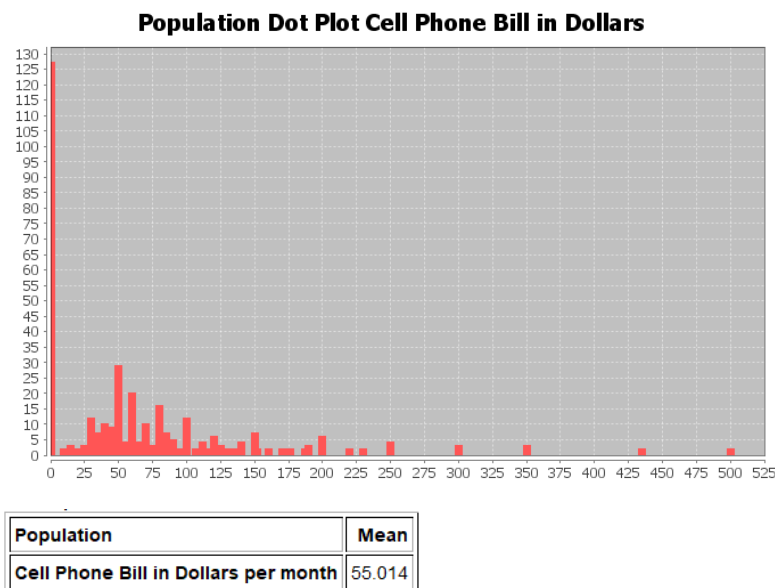
10. The following graph was created with StatKey from the “sleep hours per night” census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



- What was the shape and mean average of the population?
- If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

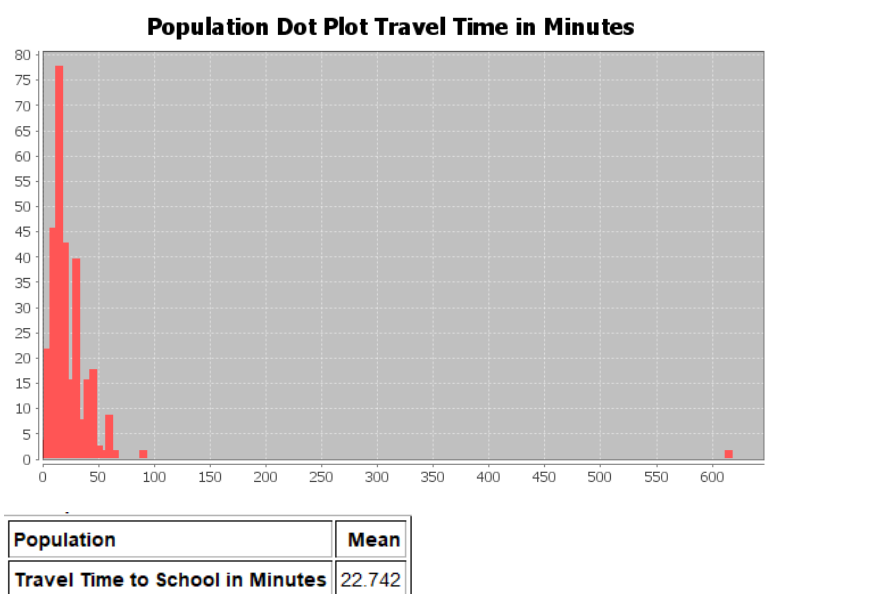


11. The following graph was created with StatKey from the cell phone bill (in dollars per month) census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



- What was the shape and mean average of the population?
- If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?

12. The following graph was created with StatKey from the travel time to school in minutes census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.

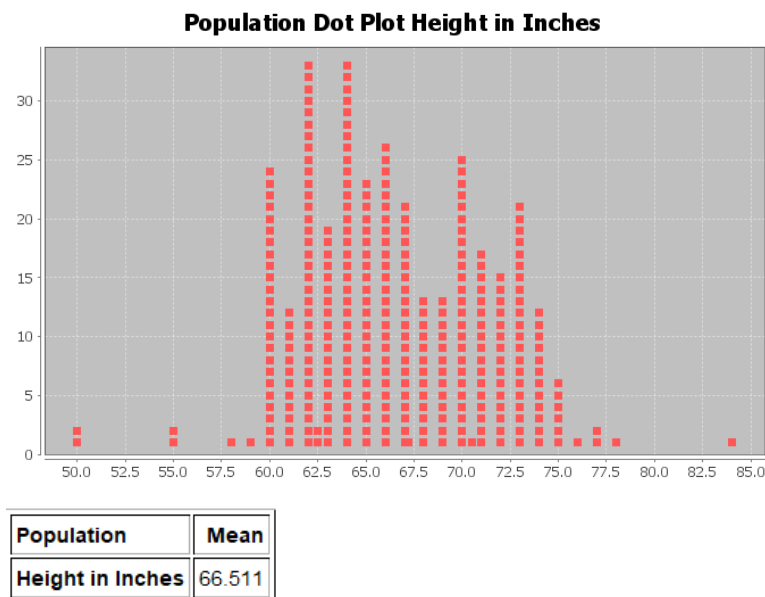


- What was the shape and mean average of the population?
- If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

13. The following graph was created with StatKey from the height in inches census data (Math 140 Survey Data). Assume this census represents the population of Math 140 statistics students at College of the Canyons in the fall 2015 semester.



- a) What was the shape and mean average of the population?
 - b) If a random sample was taken from this population, what is the minimum sample size we should have in order to have a nearly normal sampling distribution for sample means?
14. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students with brown hair is 0.537. Use this population proportion (π) to answer the following questions.
- a) Use the formula $n = \frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
 - b) Use the formula $n = \frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
 - c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?
15. A census of COC statistics students in the fall 2015 semester indicated that the population proportion of statistics students that smoke cigarettes is 0.091. Use this population proportion (π) to answer the following questions.
- a) Use the formula $n = \frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
 - b) Use the formula $n = \frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
 - c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?



16. Approximately 60% of college students in the U.S. were able to finish their bachelor's degree in six years. Use this population proportion (π) to answer the following questions.

- Use the formula $n = \frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
- Use the formula $n = \frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
- If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

17. Approximately 9.4% of all adults in the U.S. have diabetes. Use this population proportion (π) to answer the following questions.

- Use the formula $n = \frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
- Use the formula $n = \frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
- If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

18. Approximately 90% of all lung cancer cases are caused by cigarette smoking. Use this population proportion (π) to answer the following questions.

- Use the formula $n = \frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
- Use the formula $n = \frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
- If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?

19. Approximately 10% of all people are left handed. Use this population proportion (π) to answer the following questions.

- Use the formula $n = \frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
- Use the formula $n = \frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
- If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?



20. Approximately 2% of all vehicles sold in the U.S have a manual transmission. Use this population proportion (π) to answer the following questions.

- a) Use the formula $n = \frac{10}{(\pi)}$ to calculate the minimum sample size to get at least ten successes in our sample data.
 - b) Use the formula $n = \frac{10}{(1-\pi)}$ to calculate the minimum sample size to get at least ten failures in our sample data.
 - c) If our sample data has at least ten successes and at least ten failures, then we expect the sampling distribution for sample proportions to be approximately normal. What is the minimum sample size we expect to have a nearly normal sampling distribution for sample proportions?
-

Section 2D – Introduction to Confidence Intervals

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Point Estimate: When someone takes a sample statistic and then claims that it is the population parameter.

Margin of Error: Total distance that a sample statistic might be from the population parameter. For normal sampling distributions and a 95% confidence interval, the margin of error is approximately twice as large as the standard error.

Standard Error: The standard deviation of a sampling distribution. The distance that typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distributions is usually close to the population parameter, the standard error tells us how far typical sample statistics are from the population parameter.

Confidence Interval: Two numbers that we think a population parameter is in between.

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

What is the population percentage of people worldwide that have congestive heart failure (CHF)? What is the population mean average salary of every working adult in Japan? Estimating population parameters is very important if we are to understand the world around us.



Estimating Population Parameters

There are two ways for finding a population parameter, an unbiased census or the center of a sampling distribution from thousands of large random samples. If you collect data from everyone in your population, and have not incorporated bias into the data, then you have collected an unbiased census. In that case, you know the entire population. Unbiased census data can be used to find population parameters like the population mean (μ), the population standard deviation (σ), or the population proportion (π). Simply calculate the mean, proportion, or standard deviation of the census and you know your population parameter.

We also learned that if you collect many large random samples from a population, you could create a sampling distribution. The center of the sampling distribution is usually a very good estimate of the population parameter.

This is not what happens usually in the real world. Populations may have millions of people, making it virtually impossible to take a census (unless you are the government). Most data scientists simply cannot collect a census from large populations. Random samples are usually very difficult to collect and can be expensive. Therefore, it is rare to see someone collect many random samples from the same population. Certainly not thousands of random samples. Therefore, we often cannot create a sampling distribution from the population either.

A person analyzing data usually has one large random sample. The question is can we estimate a population parameter with one large random sample?

Remember the principle of sampling variability.

Sampling Variability: Random sample statistics will usually be different from each other and different from the population parameter.

Every time we take a random sample, we will get something different. The sample statistic you calculate from random sample data will usually be off from the population parameter. Remember there will always be a margin of error.

Key: If all you have is one random sample, you will not be able to find the population parameter. The sample statistic you calculate will be off from the real population parameter.

If we have one random sample, can we estimate the population parameter at least? Yes, but we should be careful how we label it.

Point Estimates

Point Estimate: Some people take the random sample statistic and then just tell us in their article or report that the sample statistic is the population parameter.

Most of the time, when someone in an article gives us a population parameter, it usually is not the actual population parameter. It is a point estimate. They took some sample data, calculated the sample mean, and then tell us that the sample mean is the population mean. As you can imagine this creates a lot of confusion. Many people read articles and think the author knows the exact population mean or the exact population percentage, when in fact the number the author is quoting came from a sample. It is important to be aware of this. A good scientific report will usually make this distinction.

Good Point Estimate: "We tested a random sample of people for high cholesterol and found that 31.7% of the sample had high cholesterol. So we estimate that the population percentage of people worldwide with high cholesterol is about 31.7% with a 1.2% margin of error."

Bad Point Estimate: "The population percentage of people worldwide that have high cholesterol is 31.7%."

The second example shows what most articles say. It can be very confusing for most people since they believe that the author knows the population percentage for everyone worldwide. They do not realize it was just a sample percentage. We know from our study of sampling distributions and sampling variability that this sample percentage can be far off from the real population percentage.



Confidence Intervals

A sample statistic will usually be off from the population parameter. In other words, the sample statistic has a margin of error.

Margin of Error: The distance that a sample statistic might be from the population parameter.

It is relatively easy to calculate margin of error if already know the population parameter. Remember we rarely know the population parameter in the real world. It can be very difficult to estimate margin of error when you do not know the population parameter. Many mathematicians and statisticians put a lot of thought into finding formulas that would estimate the margin of error. We will go over some of these famous margin of error formulas throughout the chapter.

If you know the margin of error and the sampling distribution was relatively normal or symmetric, then you can use the margin of error to create a confidence interval.

Confidence Interval: Two numbers that we think a population parameter is in between.

When all you have is one random sample, you will not be able to find the population parameter exactly, but you can find two numbers that we think the population parameter may be in between. This is called a “confidence interval”. For example, we may know what the population percentage is, but we think it is between 10.2% and 13.6%.

Here is a common formula for calculating a confidence interval.

Sample Statistic \pm Margin of Error

Example 1: Suppose we look at a random sample of gas mileage (miles per gallon) for various cars. We want to estimate the population mean average mpg for all cars in the world. The sample mean (\bar{x}) was 24.761 mpg but remember this does not mean that the population mean is 24.761. Using a formula, we were able to calculate the margin of error for this sample to be 2.152 mpg. So what would the confidence interval be?

Sample Statistic \pm Margin of Error

$\bar{x} \pm$ Margin of Error

24.761 mpg \pm 2.152 mpg

Lower Limit: 24.761 – 2.152 = 22.609 mpg

Upper Limit: 24.761 + 2.152 = 26.913 mpg

Therefore, a sample mean average gas mileage of 24.761 mpg tells us that the population mean average gas mileage for cars could be in between 22.609 mpg and 26.913 mpg.

Confidence Intervals can be written in three ways: interval notation, inequality notation, or just give the sample statistic and margin of error. In this example, here are the three ways the confidence interval may be written.

Interval Notation: (22.609 mpg , 26.913 mpg)

Most computer programs write their confidence intervals in interval notation. This does not mean (x , y) like in algebra. It means the population parameter could be any of the millions of numbers in between 22.609 mpg and 26.913 mpg.

Inequality Notation: 22.609 mpg $< \mu <$ 26.913 mpg

Remember this interval was trying to find two numbers that the population mean (μ) is in between. That is exactly what this says.

Sample Statistic and Margin of Error: 24.761 mpg (\pm 2.152 mpg error)

Many scientific journals or articles write it this way. They write the sample statistic as their point estimate with the margin of error.



Example 2: In the article earlier, we were looking for the percentage of people worldwide with high cholesterol. What would the confidence interval be for this problem?

"We tested a random sample of people for high cholesterol and found that 31.7% of the sample had high cholesterol. There was a 1.2% margin of error."

When calculating confidence intervals from a percentage, we usually convert the sample percentage into a sample proportion (\hat{p}). We should also convert the margin of error into a proportion.

$$31.7\% = 0.317$$

$$1.2\% = 0.012$$

Sample Statistic \pm Margin of Error

$\hat{p} \pm$ Margin of Error

$$0.317 \pm 0.012$$

$$\text{Lower Limit: } 0.317 - 0.012 = 0.305$$

$$\text{Upper Limit: } 0.317 + 0.012 = 0.329$$

We can convert this proportion back into percentages if we wish. Notice that we can write the confidence interval in three ways again. Remember a population proportion can be written with the letter "p" or " π ".

Interval Notation: (0.305 , 0.329) or (30.5% , 32.9%)

Inequality Notation: $0.305 < \pi < 0.329$ or $30.5\% < \pi < 32.9\%$

Sample Statistic and Margin of Error: 31.7% (\pm 1.2% error)

You should be comfortable converting percentages into proportions and proportions into percentages. Notice that when calculating the upper and lower limits we could have added and subtracted the percentages and got the same answer.

$$\text{Lower Limit: } 31.7\% - 1.2\% = 30.5\%$$

$$\text{Upper Limit: } 31.7\% + 1.2\% = 32.9\%$$

So a sample percentage of 31.7% does not tell us that the population percentage. It tells us that the population percentage could be in between 30.5% and 32.9%.

Important Note: Never add or subtract a proportion and a percentage. Yes, they are equivalent, but they are not the same. Either add and subtract the proportions, or add and subtract the percentages.

Never do this!! 11.9 ± 0.017

In the last two examples, how confident are we about these results?

Confidence Levels

When calculating confidence intervals, it is important to know what "confidence level" was used. A confidence level is not an abstract feeling about how confident you are. It is tied to the mathematical calculation of the margin of error. The most common confidence levels are 90%, 95% and 99%, with 95% being by far the most common. Whenever you ask a computer to calculate a confidence interval you must choose what level you want to use. Usually it is 95%.

Think of it this way. The more confident you are, the larger the margin of error and the wider you make the confidence interval. That way you are more likely to have the actual population parameter in between the two numbers. The less confident you are the smaller the margin of error and the narrower the confidence interval. I like to think of the confidence level as a catcher's mitt in baseball. If I want to be 90% confident that I catch the ball (catch



*This chapter is from Introduction to Statistics for Community College Students,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

the population parameter), I will use a regular sized catcher's mitt. If I want to be 95% confident I catch the ball, I will use a jumbo-sized catcher's mitt. If I want to be 99% confident that I catch the ball, I will use a huge soccer net.

90% confidence level → Small margin of error → Narrow confidence interval (*Regular sized Mitt*)

95% confidence level → Larger margin of error → Wider confidence interval (*Jumbo sized Mitt*)

99% confidence level → Extremely Large margin of error → Very wide confidence interval (*Soccer Net*)

Example: Earlier we looked at creating a confidence interval to estimate two numbers that we think the population mean average gas mileage (mpg) is in between. The following printout shows the calculation for 90%, 95% and 99% confidence levels. Notice that as the confidence level increases, the margin of errors are increasing the numbers in the confidence intervals are getting farther apart.

Confidence Interval - One population mean: confidence level = 0.9

Input: C1 MPG

σ unknown

Var	N	Mean	Stdev	Margin of Error	90.0%CI
C1 MPG	38.0	24.761	6.547	1.792	(22.9686, 26.5524)

Confidence Interval - One population mean: confidence level = 0.95

Input: C1 MPG

σ unknown

Var	N	Mean	Stdev	Margin of Error	95.0%CI
C1 MPG	38.0	24.761	6.547	2.152	(22.6085, 26.9126)

Confidence Interval - One population mean: confidence level = 0.99

Input: C1 MPG

σ unknown

Var	N	Mean	Stdev	Margin of Error	99.0%CI
C1 MPG	38.0	24.761	6.547	2.884	(21.8765, 27.6446)

Confidence Interval Sentence

Computers can calculate confidence intervals. The job of a data analyst, data scientist or statistician is to explain. In other words, the sentences are very important. Whenever we write a sentence to explain a confidence interval, we should always state the confidence level that was used. For one-population confidence intervals, we should also give the two numbers that the population parameter is in between.

One Population Confidence Interval Sentence:

"We are (90%, 95% or 99%) confident that the population parameter is in between # and #".

Here is the sentence for the 90% confidence interval estimate of the population mean average mpg. Remember in quantitative data, you can round the answers to one more decimal point than is present in the original data. (In this case, since the original sample data mpg values were rounded to the tenths place, we can round the confidence intervals to the hundredths place.) If you do not know the accuracy of your data, it is better not to round the numbers.



We are 90% confident that population mean average gas mileage for all cars is between 22.97 mpg and 26.55 mpg.

Here is the sentence for the 95% confidence interval estimate of the population mean average mpg. We rounded the Statcato answers to the hundredths place.

We are 95% confident that population mean average gas mileage for all cars is between 22.61 mpg and 27.64 mpg.

Here is the sentence for the 99% (rounded) confidence interval estimate of the population mean average mpg.

We are 99% confident that population mean average gas mileage for all cars is between 21.88 mpg and 26.91 mpg.

Here is the sentence for the genetic trait population percentage confidence interval. We will assume it was a 95% confidence level.

We are 95% confident that population percentage of all people worldwide that have high cholesterol is between 30.5% and 32.9%.

Understanding Confidence Levels

Here are the definitions of confidence. Notice that these definitions are not talking about a “feeling” about confidence. They are also not talking about being sure that the population parameter is in between the exact two numbers in a confidence interval. So what do these mean?

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

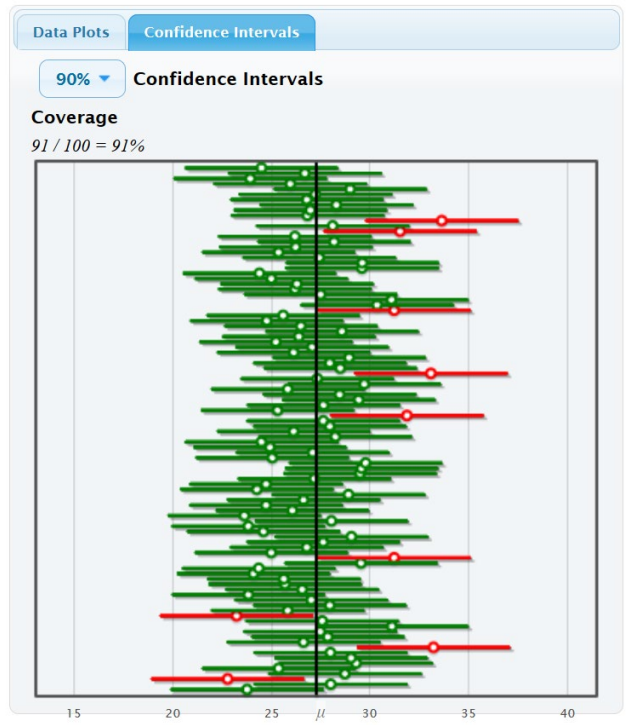
These definitions are talking about many samples, many confidence intervals. In essence, a sampling distribution.

Example 1: 90% confidence level sampling distribution

Work Hours per Week for working COC Statistics Students (Fall 2015 semester)

In a previous section, we created a sampling distribution of sample means for the work hours of statistics students. We did not use students that said they work “zero” hours. To understand confidence levels, we are going to take it a step further. Instead of just taking many random samples and calculating many sample means, we are going to use StatKey to calculate many confidence intervals. All of the confidence intervals will have a 90% confidence level. We used a sample size of 30 this time.

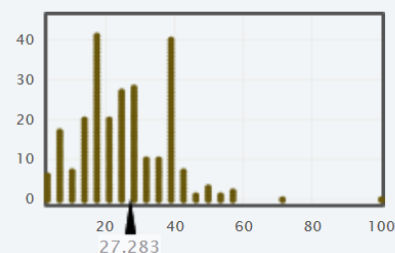




Let us see if we understand what we are looking at. The dark line indicates the population mean of 27.283 hours of work per week. If the population mean is in between the two numbers in the confidence level, then the confidence interval is green. This indicates that the confidence interval contains the population parameter. If the population mean is not in between the two numbers in the confidence level, then the confidence interval is red. This indicates that the confidence interval does not contain the population parameter.

Population

$n = 258$, mean = 27.283
median = 25, stdev = 12.969



Notice when we use a 90% confidence level, about 90% of them were green (contained the population parameter) and about 10% of them were red (did not contain the population parameter). In other words, not all confidence intervals contain the population parameter! This is what the definition of 90% confidence is talking about. If we take many random samples, and create many confidence intervals, about 90% of the confidence intervals will have the population parameter in between the two numbers and 10% of them will not.

Notice that the green and red lines describing the confidence interval have a lot of variability. This is sampling variability at work. Random samples will always be different. That means that the confidence interval numbers will also be different for every random sample.

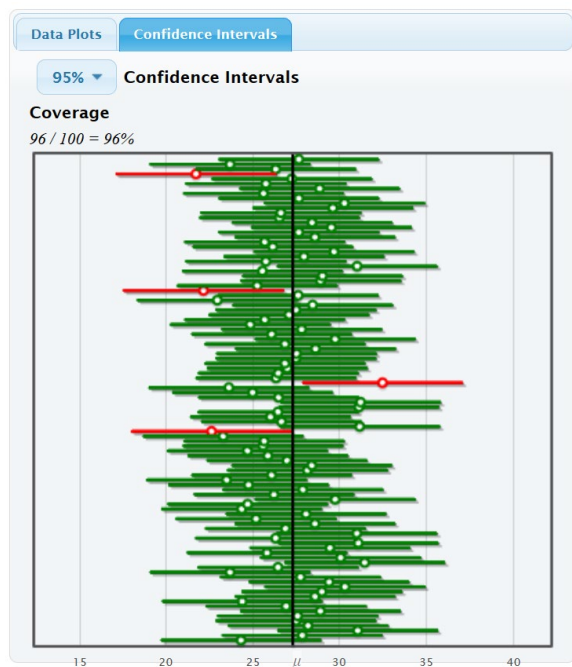


This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Also, notice that the number of green confidence intervals was not exactly 90%. In fact, it was 91% for the first hundred samples. 90% is a limit. This means that because of sampling variability, the exact percentage of green confidence intervals will fluctuate. As the number of samples increase, the number usually gets closer and closer to 90%.

Example 2: 95% confidence level sampling distribution

Work Hours per Week for working COC Statistics Students (Fall 2015 semester)

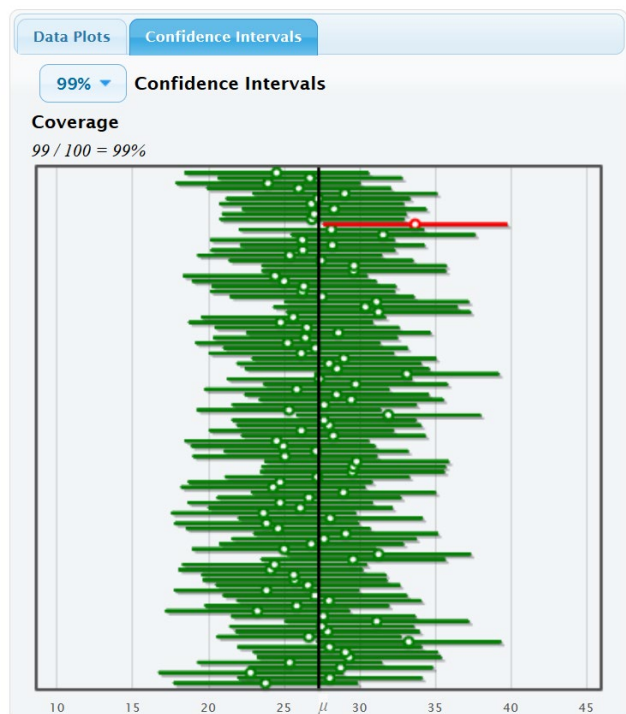


Now we will set the confidence levels to 95%. We calculated many confidence intervals and all of them have a 95% confidence level. Notice that the percentage of green confidence intervals that contain the population mean average is now approaching 95%. It is actually 96% for these first 100 samples, but as the number of samples increase, the percentage will get closer to 95%. Also, notice that the percentage of red confidence intervals that do not contain the population mean average is now approaching 5%. It is actually 4% for these first 100 samples, but as the number of samples increase, the percentage will get closer to 5%. This again is what the definition of 95% confidence is talking about. If we create many 95% confidence intervals, about 95% of them will be green and contain the population parameter, and about 5% of them will be red and not contain the population parameter.



Example 3: 99% confidence level sampling distribution

Work Hours per Week for working COC Statistics Students (Fall 2015 semester)



If we set the confidence levels to 99%, we see that the percentage of green confidence intervals that contain the population mean average is now approaching 99% and the percentage of red confidence intervals that do not contain the population mean average is now approaching 1%. This again is what the definition of 99% confidence is talking about. If we create many 99% confidence intervals, about 99% of them will be green and contain the population parameter, and about 1% of them will be red and not contain the population parameter.

Finding the sample statistic and margin of error from a confidence interval

Occasionally you may have an article or scientific report that gives a confidence interval to estimate a population mean or a population proportion, yet neglects to tell you the margin of error or the sample statistic. If you have a bootstrap distribution that looks relatively normal, you will know the confidence interval, but may not know the margin of error. Some computer programs will tell you the upper and lower limit of the confidence interval but not tell you the margin of error. In these situations, there is a way to figure out the sample statistic and the margin of error. Remember, these formulas are only used when you know the upper and lower limit of the confidence interval and you have a normal sampling distribution.

Confidence Interval Back-Solving Formula for Sample Statistic: $Sample\ Statistic = \frac{(Upper\ Limit + Lower\ Limit)}{2}$

Confidence Interval Back-Solving Formula for Margin of Error: $Margin\ of\ Error = \frac{(Upper\ Limit - Lower\ Limit)}{2}$

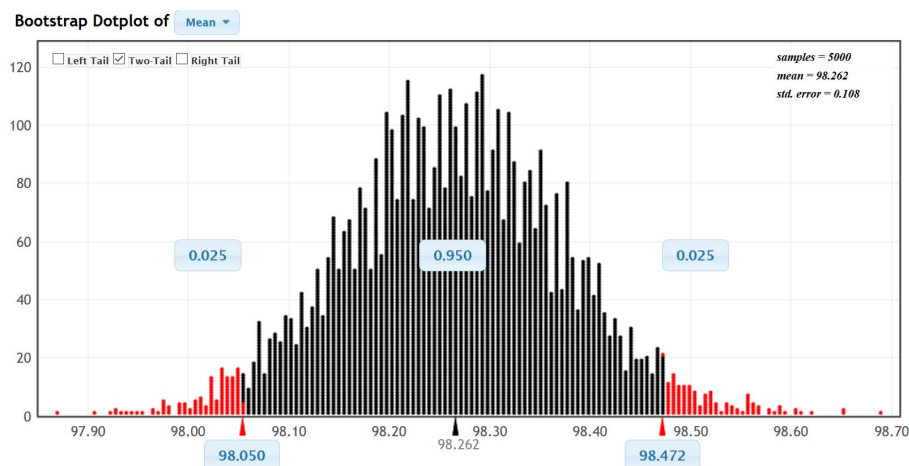
Example 1: Bootstrap Confidence Interval for Population Mean Average Body Temperature

Bootstrapping is a technique for calculating confidence intervals. The following bootstrap confidence interval was calculated from a random sample of 50 adult body temperatures in degrees Fahrenheit. The upper and lower limits for the confidence interval are given at the bottom right and left of the bootstrap distribution. The confidence level is given in the middle of the bootstrap distribution. So the 95% confidence interval is (98.050 °F, 98.472 °F).



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

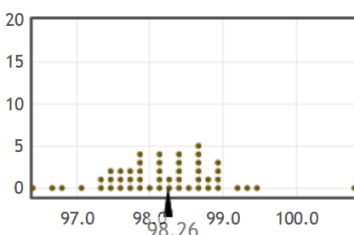
Confidence Interval Sentence: We are 95% confident that the population mean average body temperature of human adults is between 98.050°F and 98.472°F.



StatKey did tell us that the sample mean was 98.26°F, but notice that we do not know the margin of error. This is a perfect time to use the back-solving formula for margin of error.

Original Sample

$n = 50$, mean = 98.26
median = 98.2, stdev = 0.765



$$\text{Margin of Error} = \frac{(\text{Upper Limit} - \text{Lower Limit})}{2} = \frac{(98.472 - 98.050)}{2} = \frac{(0.422)}{2} = 0.211 \text{ } ^\circ\text{F}$$

Let us check the sample statistic formula and see how close it is to the actual sample mean.

$$\text{Sample Statistic (mean)} = \frac{(\text{Upper Limit} + \text{Lower Limit})}{2} = \frac{(98.472 + 98.050)}{2} = \frac{(196.522)}{2} = 98.261 \text{ } ^\circ\text{F}$$

Notice the sample statistic is very close to the actual sample mean of 98.26 °F.

Example 2: An article claims that the population percentage of young adults ages 18-25 years in the U.S. that have depression is in between 9.59% and 12.27%. This is a confidence interval. We will assume they used a 95% confidence level. What was the sample proportion and the margin of error? Again, this would be a good time to use the back-solving formulas. Remember to either use the proportions or the percentages but do not add or subtract a proportion and a percentage.

$$\text{Margin of Error} = \frac{(\text{Upper Limit} - \text{Lower Limit})}{2} = \frac{(0.1227 - 0.0959)}{2} = \frac{(0.0268)}{2} = 0.0134 \text{ (or 1.34\%)}$$

$$\text{Margin of Error} = \frac{(\text{Upper Limit} - \text{Lower Limit})}{2} = \frac{(12.27\% - 9.59\%)}{2} = \frac{(2.68\%)}{2} = 1.34\% \text{ (or 0.0134 as a proportion)}$$

$$\text{Sample Statistic (proportion)} = \frac{(\text{Upper Limit} + \text{Lower Limit})}{2} = \frac{(0.1227 + 0.0959)}{2} = \frac{(0.2186)}{2} = 0.1093 \text{ (or 10.93\%)}$$



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

$$\text{Sample Statistic (percentage)} = \frac{(\text{Upper Limit} + \text{Lower Limit})}{2} = \frac{(12.27\% + 9.59\%)}{2} = \frac{(21.86\%)}{2} = 10.93\% \text{ (or 0.1093)}$$

Summary of Confidence Intervals

- Be aware of point estimates. When a person claims to know the exact population parameter, they probably just calculated a sample statistic and are telling you it is the population parameter. We only know the population parameter if we have collected a census or if we have collected many, many random samples and look for the center of the sampling distribution. We can never know the exact population parameter from a large population if all we have is one random sample. If we have one random sample, all we can do is estimate the population parameter with a confidence interval.
- A confidence interval is two numbers that we think a population parameter is in between.
- Remember, a confidence interval should never be calculated from a census. If you already know the population parameter, there is no need to estimate it with a confidence interval. Confidence interval are calculated when we only have random sample data and need to estimate the population parameter.
- It is important to know what confidence levels were used. 90%, 95% and 99% are all sometimes used, though 95% is the most common. Remember, these levels do not refer to a feeling of confidence about one confidence interval. They are part of the confidence interval calculation and refer to the process of calculating thousands of confidence intervals.
- Here are definitions of 90%, 95%, and 99% confidence. These definitions imply that not all confidence intervals contain the population parameter. Sometimes the population parameter will not be in between the two numbers in the confidence interval.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

- The margin of error is how far we think the sample statistic is from the population parameter. A common formula that is sometimes used to calculate a confidence interval is the sample statistic \pm margin of error.
- Be able to explain the confidence interval. Here is a common sentence used for one-population confidence intervals: We are (90%, 95% or 99%) confident that the population parameter (*mean, proportion, median, standard deviation, or variance*) is in between # and #.
- If you know the upper and lower limit of a confidence interval from a normal sampling distribution, you can use these back solving formulas to find the sample statistic and the margin of error.

$$\text{Sample Statistic} = \frac{(\text{Upper Limit} + \text{Lower Limit})}{2}$$

$$\text{Margin of Error} = \frac{(\text{Upper Limit} - \text{Lower Limit})}{2}$$



Problem Set Section 2D

(For #1-10) Add and subtract the given sample statistic and margin of error to find the confidence interval estimate of the population value. Then write the confidence interval using both inequality notation and using interval notation. Now write a sentence explaining the confidence interval to someone.

Confidence interval = Sample Statistic \pm Margin of Error

1. "What is the population percent of the adult population is infected with this disease?"
Sample percentage = 4.9%
Margin of error = 1.3% (Found with 95% confidence level.)
2. "What is the population mean average height for men?"
Sample mean = 68.335 inches
Margin of error = 1.293 inches (Found with 99% confidence level.)
3. What is the population standard deviation for the systolic blood pressure in women?
(Assume there was a normal sampling distribution.)
Sample standard deviation = 17.11 mm of Hg
Margin of error = 3.31 mm of Hg (Found with 90% confidence level.)
4. What is the population percentage of left-handed people get migraine headaches?
Sample proportion = 0.088
Margin of error = 0.027 (Found with 95% confidence level.)
5. What is the population mean average price of a used mustang car in thousands of dollars?
Sample mean = 15.98 thousand dollars
Margin of error = 3.78 thousand dollars (Found with 90% confidence level.)
6. "What is the population percentage of rabid animals are wild?"
Sample proportion = 0.903
Margin of error = 0.008 (Found with 95% confidence level.)
7. "What is the population mean average weight for men?"
Sample mean = 172.55 pounds
Margin of error = 11.272 pounds (Found with 99% confidence level.)
8. What is the population variance for the heights of men? (Assume there was a normal sampling distribution.)
Sample variance = 10.177 square inches
Margin of error = 3.661 square inches (Found with 90% confidence level.)
9. What is the population percentage of women in the U.S. are overweight?
Sample percentage = 36.9%
Margin of error = 1.44% (Found with 95% confidence level.)
10. What is the population mean average amount of tip in dollars at a particular restaurant?
Sample mean = \$3.849
Margin of error = \$0.504 (Found with 99% confidence level.)

(For #11-20) Write a sentence explaining each of the following confidence intervals. Then use the following formulas to identify the sample statistic (\hat{p} or \bar{x} or s) and the margin of error.

$$\text{Sample Statistic} = \frac{(\text{upper limit} + \text{lower limit})}{2} \quad \text{Margin of Error} = \frac{(\text{upper limit} - \text{lower limit})}{2}$$

11. A 95% confidence interval estimate of the population proportion of fat in the milk from Jersey cows is (0.046 , 0.052).
12. A 99% confidence interval estimate of the population mean number of miles is $13.4 \text{ miles} < \mu < 17.2 \text{ miles}$.



13. A 90% confidence interval estimate of the population proportion of people who will vote for the Independent party candidate is $0.068 < \pi < 0.083$.
14. A 95% confidence interval estimate of the population mean amount of milk in gallons is $(48.7, 58.4)$.
15. A 99% confidence interval estimate of the population standard deviation for the height of men in inches is $2.34 < \sigma < 2.87$. Assume there was a normal sampling distribution.
16. A 95% confidence interval estimate of the population proportion of peanuts in a can of mixed nuts is $0.4221 < \pi < 0.6179$.
17. A 99% confidence interval estimate of the population mean pH of lakes in Florida is $(6.118, 7.064)$.
18. A 90% confidence interval estimate of the population proportion of home teams that win a soccer game is $(0.5093, 0.6574)$.
19. A 95% confidence interval estimate of the population mean average price of apartments in Manhattan, NY is $\$2514.36 < \mu < \3798.64 .
20. A 90% confidence interval estimate of the population variance for the pH of lakes in Florida is $1.2353 < \sigma^2 < 2.3675$. Assume there was a normal sampling distribution.

(#21-26) Go to www.matt-teachout.org and click on “statistics” and then “data sets”. Open the “coffee data” and copy and Columbian Mild price data. Now go to www.lock5stat.com and click on the “StatKey” button. Under the “sampling distribution” menu click on “mean”. Under edit data, paste the Columbian Mild coffee data. Click on “samples of size n” and put in 30. Turn off the button that says, “First column is identifier” as we have only a single column of data. Now click ok. You are now ready to create your sampling distribution. This time we want the computer to create a confidence interval for each sample it takes. On the right side of the screen, click on the button that says confidence intervals and set the confidence level to 95%. StatKey will take a random sample from the population data, find the sample mean and place a dot for the sample mean in the distribution. It will also create a confidence interval from that sample mean. StatKey will keep track of whether the true population mean is actually contained in the confidence interval or not. Green means the confidence interval did contain the population value and red means that the confidence interval did not contain the population value. Now answer the following questions.

21. Notice the confidence intervals for sample means were different for each random sample. Discuss the implications of sampling variability on the accuracy of a confidence interval from a random sample.
22. What was the population mean in cents per pound? Did all the confidence intervals contain the population mean? What does it mean that the interval “contained” or “captured” the population mean?
23. How many total random samples did you take? How many of them contained the population mean? What percent of the confidence intervals contained the population mean?
24. How many confidence intervals did not contain the population mean? What percent of the confidence intervals did not contain the population mean?
25. As the number of random samples increased, did the percentage of confidence intervals that contained the population mean get closer or farther away from 95%? Why do you think that is?
26. Here is the definition of 95% confidence: “95% of confidence intervals contain the population parameter and 5% do not contain the population parameter”. Explain this definition of 95% confidence in your own words.

(#27-32) Assume a fair coin has a 50% (0.5) chance of getting tails. If we take samples from that population, the sample proportions will usually not be 0.5. We want to look at lots of proportion confidence intervals from sample proportions. Go to www.lock5stat.com and click on the “StatKey” button. Under the “sampling distribution” menu click on “proportion”. Under “edit proportion”, put in 0.5 and then click ok. Under “sample size”, set it to “n = 30”. You are now ready to create your sampling distribution. We want the computer to create a confidence interval for each sample proportion. On the right side of the screen, click on the button that says confidence intervals and set the



confidence level to 90%. StatKey will take a random sample from the population data, find the sample proportion and place a dot for the sample proportion in the distribution. It will also create a confidence interval from that sample. Remember that the population proportion for a fair coin is 0.5 (50%). StatKey will keep track of whether the true population proportion is actually contained in the confidence interval or not. Green means the confidence interval did contain the population value and red means that the confidence interval did not contain the population value. Now answer the following questions.

27. Notice the confidence intervals for sample proportions were different for each random sample. Discuss the implications of sampling variability on the accuracy of a confidence interval created from a random sample proportion.

28. Did all the confidence intervals contain the population proportion of 0.5? What does it mean that the interval “contained” or “captured” the population parameter?

29. How many total confidence intervals did you make? How many of them contained the population proportion 0.5? What percent of the confidence intervals contained the population proportion 0.5?

30. How many of the confidence intervals did not contain the population proportion 0.5? What percent of the confidence intervals did not contain the population proportion 0.5?

31. As the number of random samples increased, did the percentage of confidence intervals that contained the population proportion get closer or farther away from 90%? Why do you think that is?

32. Here is the definition of 90% confidence: “90% of confidence intervals contain the population parameter and 10% do not contain the population parameter”. Explain this definition of 90% confidence in your own words.

Section 2E – One Population Mean & Proportion Confidence Intervals

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Sampling Distribution: Take many random samples from a population, calculate a sample statistic like a mean or percent from each sample and graph all of the sample statistics on the same graph.
The center of the sampling distribution is a good estimate of the population parameter.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.

Point Estimate: When someone takes a sample statistic and then claims that it is the population parameter.

Margin of Error: Total distance that a sample statistic might be from the population parameter. For normal sampling distributions and a 95% confidence interval, the margin of error is approximately twice as large as the standard error.

Standard Error: The standard deviation of a sampling distribution. The distance that typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distributions is



usually close to the population parameter, the standard error tells us how far typical sample statistics are from the population parameter.

Confidence Interval: Two numbers that we think a population parameter is in between. Can be calculated by either a bootstrap distribution or by adding and subtracting the sample statistic and the margin of error.

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval. The center of the bootstrap distribution is the original real sample statistic.

In the last section, we saw that if we have only one random sample from a population, we would not be able to find the population parameter exactly. The best we can do is create a confidence interval, which is two numbers that we think the population parameter is in between.

In this section, we will look at some of the famous formulas that statisticians use to estimate population parameters with confidence intervals. We will also look at sample data conditions in order to ensure the accuracy of the formula.

If our sampling distribution is normal, most one-population confidence interval formulas start from the following.

Sample Statistic \pm Margin of Error

Early mathematicians and statisticians thought a lot about how to estimate the margin of error when you do not know the population parameter. The key was the sampling distribution. If a sampling distribution looked normal, then the empirical rule would suggest that the middle 95% would correspond to two standard deviations above and below the center. This gave rise to another common formula.

Sample Statistic $\pm (2 \times \text{Standard Error})$

Critical Value Z-scores

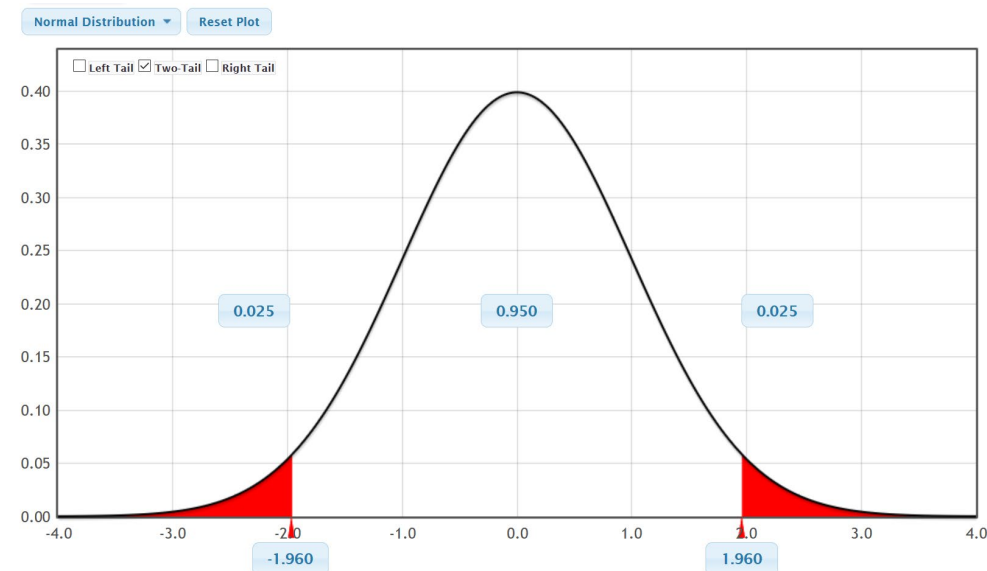
What is the “2” representing in the following formula. It seems it is counting how many standard errors one is from the mean (center) of the sampling distribution. Does this remind you of a statistic we previously learned?

If you recall, the Z-score measures the number of standard deviations from the mean. So the “2” is really a Z-score. This gave rise to the idea of replacing the “2” with a Z-score. The Z-score can be adapted for 90%, 95% or 99%. Remember two standard deviation is just an approximation for 95%. If that is the case, can we get a more accurate number for 95%?

Using a normal calculator, we can calculate the Z-score for 90%, 95% and 99% confidence. These are very famous and are often referred to as “critical value Z-scores” or “ Z_c ” for short.

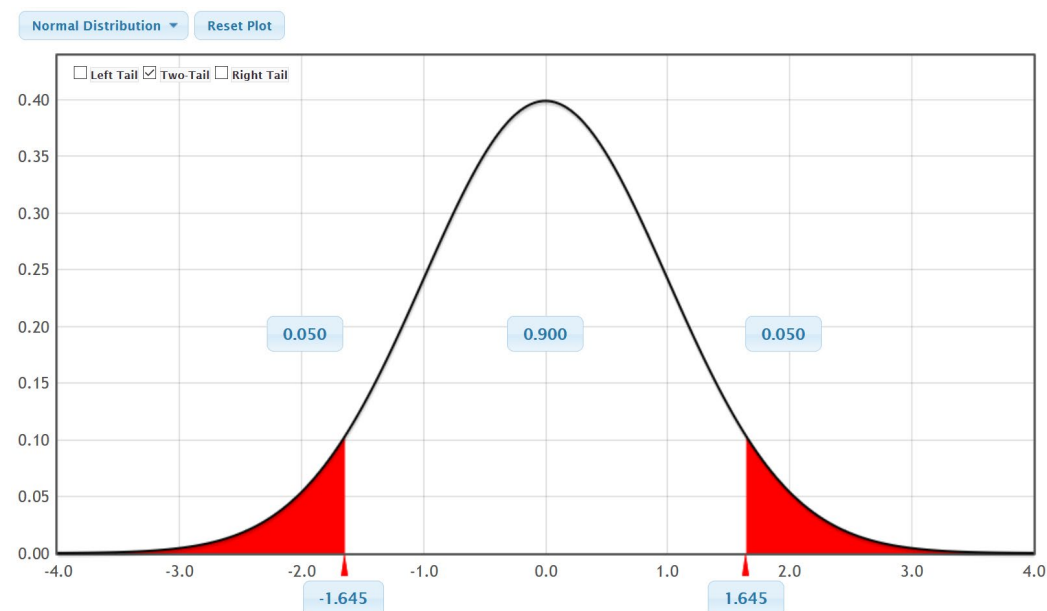


Go to www.lock5stat.com and open StatKey. Under the “theoretical distributions” menu, click on “normal”. If the mean is zero and the standard deviation is one, then this will calculate Z-scores. Click the “two-tail” button. The first Z-score calculated is for 95%.



This is the most famous of all the critical value Z-scores. Remember, for the middle 95%, the empirical rule indicates that it will be “about” two standard deviations. It turns out, 1.96 standard deviations is more accurate. Notice that just like the confidence intervals have an upper limit and lower limit, so does the Z-score critical values. For 95% confidence, we can replace the ± 2 in the formula with ± 1.96 .

What about 90% confidence intervals? Go back to the normal calculator in StatKey and click on the “0.95” in the middle. Change it to 0.9 (90%).

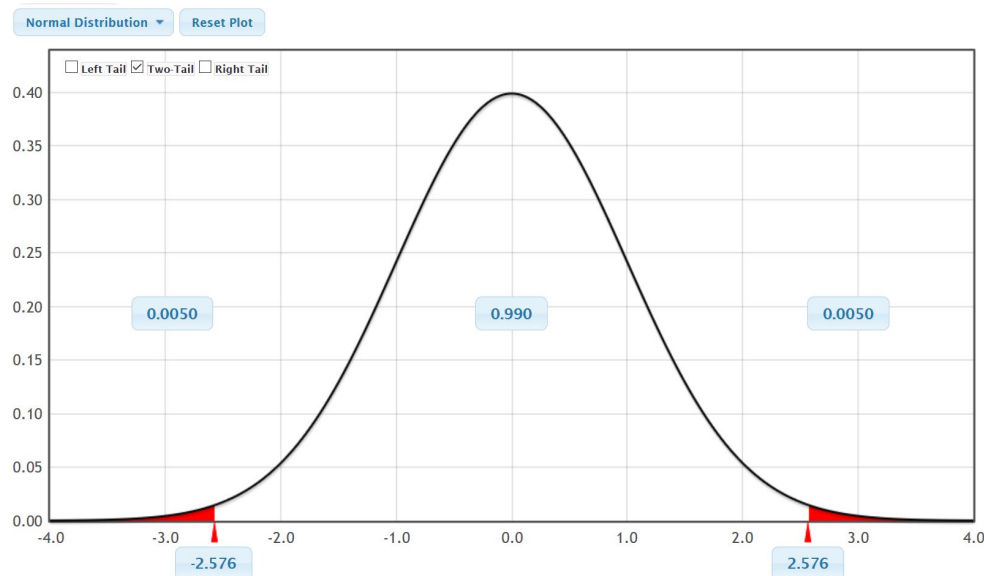


Notice the Z-score for 90% confidence intervals is ± 1.645 . Notice that as the confidence interval decreases from 95% to 90%, the Z-score gets lower. This will cause the margin of error to decrease and the confidence interval to get narrower.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

What about 99% confidence intervals? Go back to the normal calculator in StatKey and change the middle proportion into 0.99 (99%).



Notice the Z-score for 99% confidence intervals is ± 2.576 . Therefore, instead of being 1.645 standard errors away or 1.96 standard errors away, now we are 2.576 standard errors away. As the confidence interval increases from 95% to 99%, the Z-score gets larger. This will cause the margin of error to increase and the confidence interval to get wider.

Here are the famous critical value Z-scores.

- 90% confidence level: $Z = \pm 1.645$
- 95% confidence level: $Z = \pm 1.96$
- 99% confidence level: $Z = \pm 2.576$

Let us summarize the progress of our one-population confidence interval formula. It is important to remember that these formulas only work if our sampling distribution looks normal. Z-scores calculate the number of standard deviations (standard errors) from the mean in a perfectly normal curve.

Sample Statistic \pm Margin of Error

Sample Statistic $\pm (2 \times \text{Standard Error})$

Sample Statistic $\pm (Z \times \text{Standard Error})$

Statisticians discovered that as long as the sampling distribution was normal, the Z-scores were accurate for proportion (percentage) confidence intervals. The famous critical value Z-scores are still used to this day to calculate a confidence interval estimate of a population proportion (percentage).

One-Population Proportion Confidence Interval

Before computers were invented, it was very difficult to make sampling distributions. Yet it was vital to understanding sample statistics and calculating standard error. Early mathematicians and statisticians invented formulas to estimate the standard error.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

$$\text{Standard Error Estimation Formula for Proportions} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sample Proportion = \hat{p}

Sample Size = n

So now, we can finish our estimation formula for a confidence interval estimate of the population proportion. In order to estimate the margin of error, we multiply the standard error by the number of standard errors (Z-score).

Sample Statistic \pm Margin of Error

Sample Statistic \pm (2 \times Standard Error)

Sample Statistic \pm (Z \times Standard Error)

$$\hat{p} \pm \left(Z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Example 1: Calculating the confidence interval for a proportion

A random sample of 54 bears in a region of California showed that 19 of them were female. Find the sample proportion and use the formula above to calculate a 95% confidence interval estimate for the population proportion of female bears in this region of California.

$$\text{Sample Proportion } (\hat{p}) = \frac{\text{Amount of Female Bears}}{\text{Sample Size}} = \frac{19}{54} \approx 0.352$$

Critical Value Z-score for 90% Confidence = ± 1.96

Now we will replace the Z-score with 1.96 and \hat{p} with 0.352 and n with 54 into our formula and work it out. Remember to follow order of operations. Notice the standard error estimate is 0.065 (6.5%) and the margin of error estimate is 0.127 (12.7%).

$$\hat{p} \pm \left(Z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$0.352 \pm \left(1.96 \sqrt{\frac{0.352(1-0.352)}{54}} \right)$$

$$0.352 \pm \left(1.96 \sqrt{\frac{0.352(0.648)}{54}} \right)$$

$$0.352 \pm (1.96 \times 0.065)$$

$$0.352 \pm (0.127)$$

$$0.352 - 0.127 < \text{Population Proportion of Female Bears } (\pi) < 0.352 + 0.127$$

$$0.225 < \text{Population Proportion of Female Bears } (\pi) < 0.479$$

We are 95% confident that between 22.5% and 47.9% of all bears in this region of California are female.

Note: While it is important to understand formulas, data scientist today rely on computers to calculate confidence intervals. It is very difficult to calculate confidence intervals from large data sets with a formula and a calculator. The job of a data scientist, statistician, or data analyst is understand and explain the data, not to spend hours calculating something a computer can do in a split second.



To calculate this confidence interval with Statcato, we will click on the “statistics” menu and then “confidence intervals”. Click on one-population proportion and under summary data; enter 19 for the number of events and 54 for the number of trials. Set the confidence level to 0.95 and click OK.

Confidence Interval: One Population P... X

Help F1

Inputs

☐ Samples in column:

☒ Summarized sample data:

Number of events: 19

Number of trials: 54

Confidence

Confidence level: 0.95 0 - 1.00 (e.g. 0.95)

OK Cancel

Here is the Statcato printout. Notice the computer calculation is almost the same as the one we did with the formula and calculator. However, it took a lot less time.

Confidence Interval - One population proportion: confidence level = 0.95

Input: Summary data

Number of trials	Number of Events	Sample proportion	Margin of Error	95.0%CI
54	19	0.352	0.127	(0.2245, 0.4792)

Key Question: How accurate is this confidence interval?

This confidence interval relies on a Z-score and the standard error so the sampling distribution for sample proportions must be normal for this formula to be accurate. If we look at the section on the central limit theorem, we remember that for a sampling distribution for random sample proportions to be normal, we need at least ten success and at least ten failures. This gives rise to the assumptions or conditions required for certain confidence interval calculations. For the formula approach to be accurate, the following must be true. If any of these assumptions are not met, then the confidence interval may not be accurate.

One-population Proportion Assumptions

1. The categorical sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.
3. There should be at least ten successes and at least ten failures.

Let us check these assumptions in the previous confidence interval for the proportion of female bears.

1. Random Categorical Data? *Yes. This data was random and gender is a categorical variable.*
2. Data values within the sample should be independent of each other. *This can be difficult to determine. It should not be the same bear measured multiple times. In addition, if one bear is female it should not change the probability of other bears being female. It is likely safe to assume these are true in this case.*



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

3. At least ten successes? *Yes. There were 19 female bears in the data, which is more than ten.*
 At least ten failures? *Yes. There were $54 - 19 = 35$ bears that were not female in the data which is more than ten.*

Overall, it appears the data does satisfy the requirements for using the formula and so the confidence interval will be relatively accurate.

Bootstrapping

Is there a way to make a confidence interval if the data did not meet the assumptions?

It depends on which assumptions. One technique that is sometimes used is called “Bootstrapping”. Bootstrapping does require the sample to be representative of the population. That usually means it was collected randomly with data values that are independent of each other. As long as you have those two assumptions, you can bootstrap.

One-population Bootstrap Assumptions

1. The sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.

Bootstrapping does not use formula for standard error and critical values like Z-scores or T-scores. It calculates the middle 95%, 99% or 90% directly using a bootstrap sampling distribution. Since bootstrapping is not tied to formulas and critical values, it does not require the sampling distribution to be normal or to match up with a specific theoretical curve.

The idea of bootstrapping is to create a theoretical population by assuming that the population is just many copies of your one real representative random sample. In practice, bootstrapping uses computers to take thousands for random samples with replacement from your one representative random sample. It randomly selects a value from your data, but puts the value back before picking another value randomly. This allows us to get the same value in a bootstrap sample multiple times. It then calculates the statistic like the mean or proportion from all of the bootstrap samples. These are sometimes called “bootstrap statistics”. Putting all the bootstrap statistics on the same graph gives a “bootstrap sampling distribution”. If you find the computer find the cutoffs for the middle 95% of the bootstrap distribution, you have an estimated 95% confidence interval.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

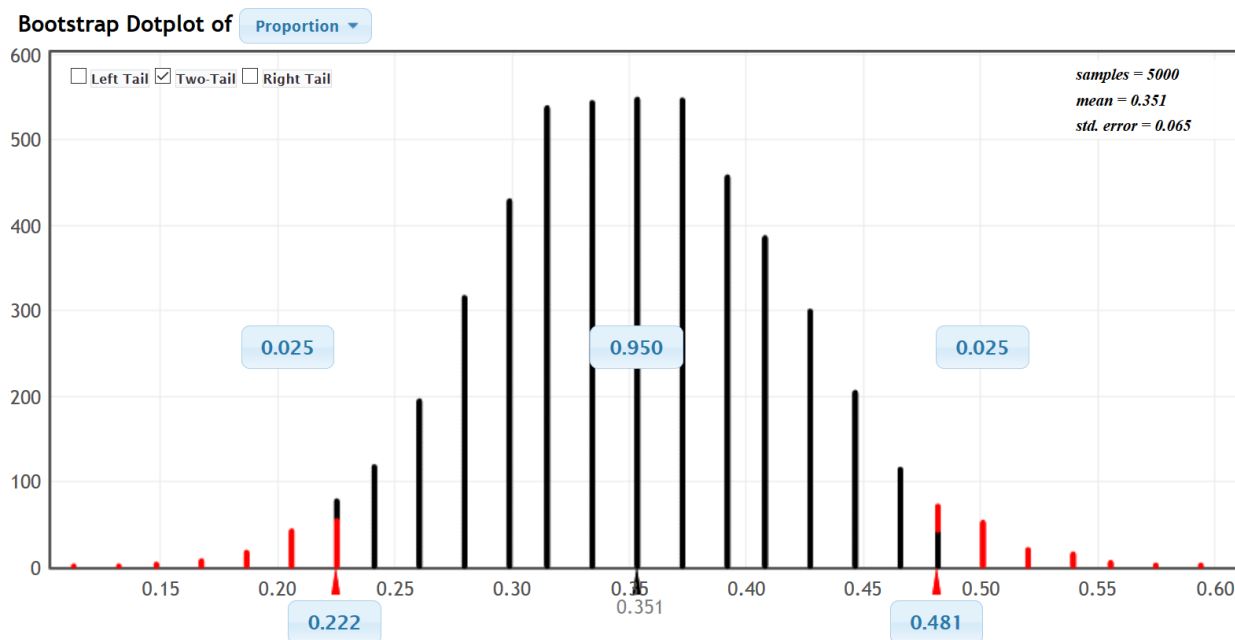
Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval.
 The center of the bootstrap distribution is the original real sample statistic.

Female Bears Example

In the last example, we used the traditional Z critical value and standard error formula to create a confidence interval and estimate the population percentage of bears that are female. We could also use a bootstrap. Go to the “Bootstrap Confidence Interval” menu in StatKey at www.lock5stat.com and click on “CI for Single Proportion”. Under “Edit Data” put in the random sample data count (19 female bears) and the total sample size (54 bears). Click “Generate 1000 Samples” a few times. Now click “Two-Tail”. The default is 95%, but you can always change the



middle proportion to 99% (0.99) or 90% (0.90) if needed. This problem was a 95% confidence interval, so we will leave the middle proportion as 0.95.



In a bootstrap confidence interval, the upper and lower limit of the confidence interval are found at the bottom right and left (0.222 and 0.481). Using these numbers, we are 95% confident that the population percentage of bears in this region of California that are female is between 22.2% and 48.1%. Notice that the upper limit, lower limit and standard error are very close to what we got by formula or Statcato. Notice that the shape of the bootstrap distribution is very normal. Though the bootstrap does not give us the margin of error like Statcato, we can use the formula we learned in the previous section. Remember the standard error and margin of error in this calculation are only reasonably accurate if the distribution is normal. Notice the margin of error is close to what we got by formula or Statcato.

$$\text{Margin of Error} = \frac{(\text{Upper Limit} - \text{Lower Limit})}{2} = \frac{(0.481 - 0.222)}{2} \approx 0.1295$$

Key Notes about Bootstrapping

- A bootstrap distribution attempts to estimate and visualize the sampling variability in the population by creating a simulated population. Remember that standard error and margin of error are only accurate if the distribution is normal. So while we can estimate standard error and margin of error from a bootstrap, they may not be accurate if the bootstrap distribution is not normal.
- While a bootstrap distribution may be similar to a true sampling distribution from the population, there are important differences. The center of a bootstrap distribution is the sample statistic from the original real random data set. This makes the bootstrap ideal for estimating the confidence interval. A true sampling distribution is taking thousands of real samples from the population, so the center of a sampling distribution is the population parameter. We should not treat a true sampling distribution from the population the same as a bootstrap. If you have a sampling distribution, then the center can get a very accurate estimate of the population parameter. If you know the population parameter, you do not need a confidence interval. The middle 95% of a sampling distribution from an actual population is not a confidence interval.



Critical Value T-scores

In 1908, a statistician named William Gosset discovered that while Z-scores were very accurate for proportions, they were not very accurate when estimating mean averages, especially if the sample size was small. Small samples should have a larger margin of error than those indicated by Z-scores. To deal with this problem, he invented T-scores. His idea was that each sample size should have a different number of standard deviations. When Gosset invented the T-distribution, he worked for Guinness Beer and was not allowed to publish his work. He therefore published under the pseudonym “student”. To this day, the T-distribution is often called the “Student T-Distribution” since it was invented by a then unknown author named “student”.

T-scores are the same as Z-scores in the sense that they count the number of standard deviations or standard errors from the mean. However, they have a built in error correction for smaller data sets. For very large sample sizes, T-scores and Z-scores are about the same. For example, if we are using a 95% confidence level and our sample size is very large, then the T-score will be close to the Z-score of ± 1.96 standard deviations. When sample sizes are small, the T-scores become significantly greater than the Z-scores. This causes the margin of error to increase for small sample sizes. Remember, less random data should result in more error. We usually use Z-scores when estimating population proportions or percentages. We prefer to use T-scores when estimating population mean averages.

Note: You can use Z-scores for the mean if the sample size is large or if you know the population standard deviation exactly. However, we rarely know the population standard deviation with any certainty, especially when we do not even know the population mean. Also in large sample sizes, the T-scores are still accurate, so you might as well use the T-scores.

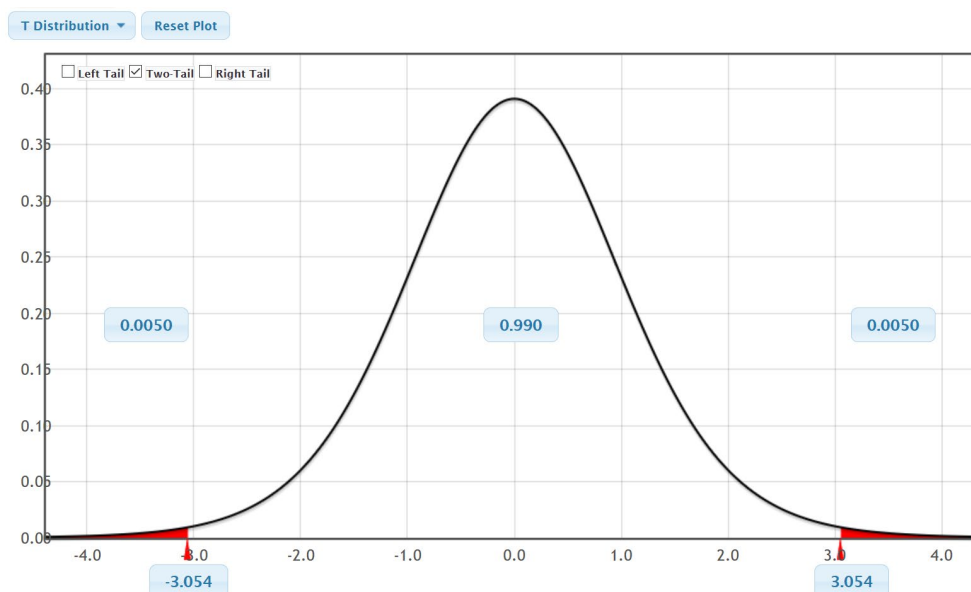
Degrees of Freedom

If you recall from previous sections, statistics like variance and standard deviation are based on a sum of squares divided by the degrees of freedom. For one sample, the degrees of freedom is usually equal to one less than the sample size ($df = n - 1$). Because of this, Gosset organized his T-scores not by sample size, but by degrees of freedom. Gosset calculated his T-scores with calculus and wrote them on charts. Before computers were invented, a statistician would first calculate the degrees of freedom and then look up the correct T-score on these charts. In modern times, T-scores can be easily calculated with computer programs like StatKey.

Example 1: Calculate the T-score critical value for a sample size $n = 13$ and a 99% confidence level.

Go to www.lock5stat.com and click on “StatKey”. Under the “theoretical distributions” menu, click on “t”. Since the sample size is 13, the degrees of freedom will be $df = 13 - 1 = 12$. If we click on “two tail” and set the middle proportion to 0.99, we will get the following.

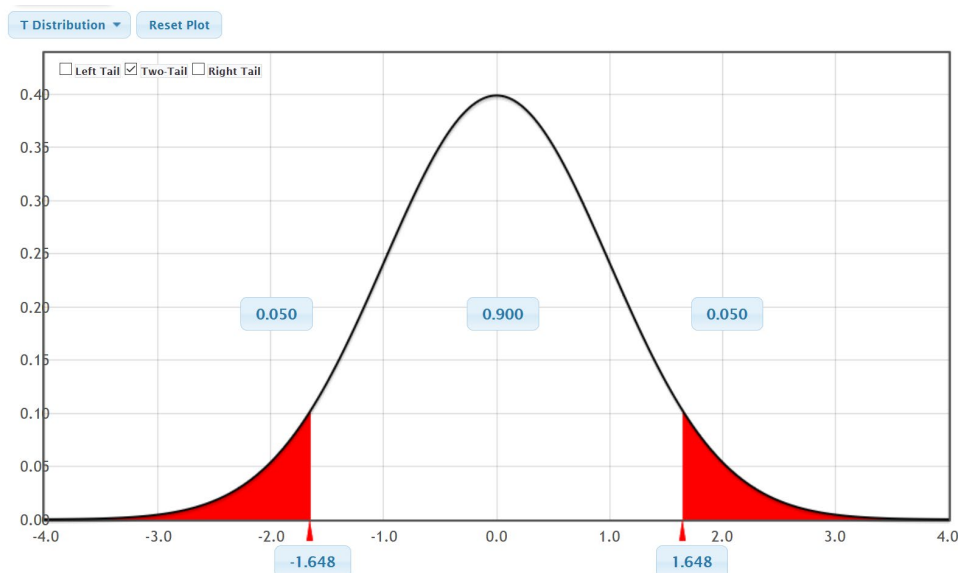




We see from the graph that critical value T-score for 99% confidence and 12 degrees of freedom is ± 3.054 . Notice this is larger than the 99% confidence critical value Z-score (± 2.576). For smaller sample sizes, the T-scores are significantly greater than the Z-scores.

Example 2: Calculate the T-score critical value for a sample size $n = 500$ and a 90% confidence level.

Go to www.lock5stat.com and click on “StatKey”. Under the “theoretical distributions” menu, click on “t”. Since the sample size is 13, the degrees of freedom will be $df = 500 - 1 = 499$. If we click on “two tail” and set the middle proportion to 0.9, we will get the following.



We see from the graph that critical value T-score for 90% confidence and 499 degrees of freedom is ± 1.648 . Notice this is very close to the 90% confidence critical value Z-score (± 1.645). For larger sample sizes, the T-scores and the Z-scores are about the same.



Summary of Critical Value T-scores

- T-scores (like Z-scores) count the number of standard deviations from the mean. In a sampling distribution of sample means, it counts how many standard errors we should be from the center of the sampling distribution for a given confidence level.
- T-scores are different for every sample size. They are usually organized by degrees of freedom. For one-population, the degrees of freedom is usually $df = n - 1$.
- T-scores are significantly larger than Z-scores for small sample sizes. The smaller the sample size, the larger the discrepancy between the T-score and Z-score.
- T-scores are about the same as Z-scores for large sample sizes.

One-Population Mean Confidence Interval

Let us look at the formula for calculating a one-population mean average confidence interval. Many computer programs to this day still use this formula.

Statisticians estimated the standard error for a sampling distribution for sample means with the following formula. The formula is surprisingly accurate and close to the standard error in an actual sampling distribution.

$$\text{Standard Error Estimation Formula for Means} = \frac{s}{\sqrt{n}}$$

Sample Standard Deviation = s

Sample Size = n

Here is the formula for a confidence interval estimate of the population mean. In order to estimate the margin of error, we multiply the standard error by the number of standard errors (T-score).

Sample Statistic \pm Margin of Error

Sample Mean \pm ($T \times$ Standard Error)

$$\bar{x} \pm \left(T \frac{s}{\sqrt{n}} \right)$$

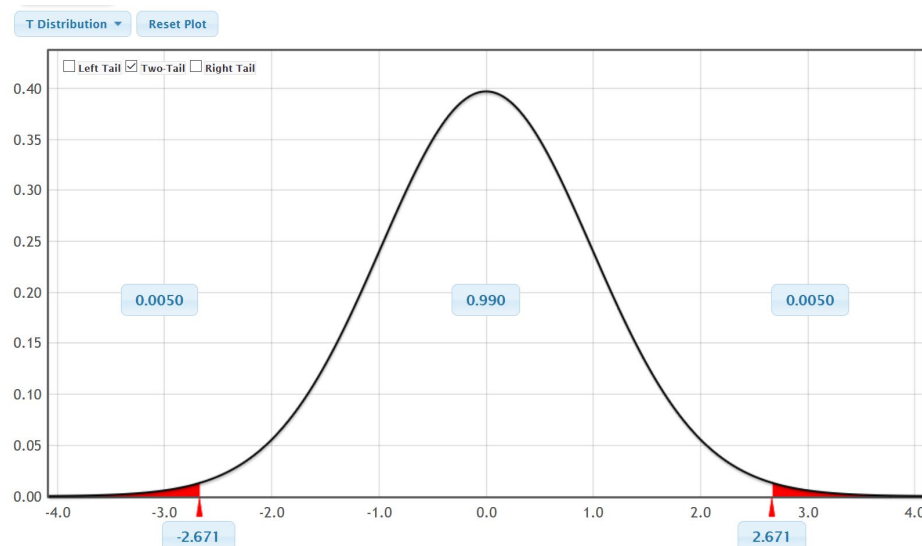
Example 1: Calculating the confidence interval estimate of a population mean

A random sample of 54 bears in a region of California was taken. The weights of the bears showed a skewed right shape with a sample mean of 182.889 pounds and sample standard deviation of 121.801 pounds. Find the degrees of freedom and the critical value T-score. Then use the formula above to calculate a 99% confidence interval estimate for the population mean average weight of bears in this region of California.

Degrees of Freedom: $df = n - 1 = 54 - 1 = 53$.

Using the T-score calculator in StatKey we found that the critical Value T-score for 99% Confidence and 53 degrees of freedom is $T = \pm 2.671$





Now we will replace the T-score with 2.671, \bar{x} with 182.889, n with 54, and s with 121.801 into our formula and work it out. Remember to follow order of operations. Notice the standard error estimate is 16.575 pounds and the margin of error estimate is 44.272 pounds.

$$\bar{x} \pm \left(T \frac{s}{\sqrt{n}} \right)$$

$$182.889 \pm 2.671 \times \frac{121.801}{\sqrt{54}}$$

$$182.889 \pm (2.671 \times 16.575)$$

$$182.889 \pm (44.272)$$

$$182.889 - 44.272 < \text{Population Mean Average Weight of Bears in Pounds } (\mu) < 182.889 + 44.272$$

$$138.617 \text{ pounds} < \text{Population Mean Average Weight of Bears in Pounds } (\mu) < 272.161 \text{ pounds}$$

We are 95% confident that the population mean average weight of bears in this region of California is in between 138.617 pounds and 272.161 pounds.

Note: While it is important to understand this formula, it is much easier to calculate this with a computer.

To calculate this confidence interval with Statcato, we will click on the “statistics” menu and then “confidence intervals”. Click on “One-population mean”. Under “Summary data”, enter 182.889 for the mean, 121.801 for the standard deviation, and 54 for the number of trials. Set the confidence level to 0.99 and click OK. If we have the raw data, we could also put in the column “C1” where it says “samples in column”.



Confidence Interval: One Population Mean

Help F1

Inputs

☐ Samples in column:
 n
 names separated by space.
 For a continuous range of columns, separate using dash (e.g. C1-C30).

☒ Summarized sample data:

Size:

Mean:

Standard deviation:

Population Standard Deviation

Population standard deviation:

☐ Known:

☒ Unknown

Confidence

Confidence level: 0 - 1.00 (e.g. 0.95)

Here is the Statcato printout. Notice the computer calculation is not exactly the same as the one we did with the formula and calculator. The computer did not round as much as we did. Computer calculations are usually much more accurate than calculator calculations because they tend to keep a lot more decimal places.

Confidence Interval - One population mean: confidence level = 0.99

Input: Summary data

σ unknown

Var	N	Mean	Stdev	Margin of Error	99.0%CI
summary	54.0	182.889	121.801	44.285	(138.6039, 227.1741)

It might be good to adjust our explanation sentence with the more accurate numbers from the computer.

We are 95% confident that the population mean average weight of bears in this region of California is in between 138.604 pounds and 272.174 pounds.

Key Question: How accurate is this confidence interval?

This confidence interval relies on a T-score and standard error so the sampling distribution for sample means must be normal for this formula to be accurate. If we look at the section on the central limit theorem, we remember that for a sampling distribution for random sample means to be normal, we need one of two things to be true. Either the data itself must be normal or the sample size must be at least 30. This gives rise to the assumptions or conditions required for mean average confidence interval calculations. For the formula approach to be accurate, the following must be true. If any of these assumptions are not met, then the confidence interval may not be accurate.

One-population Mean Assumptions

1. The quantitative sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.
3. The sample size should be at least 30 or have a nearly normal shape.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Let us check these assumptions in the previous confidence interval for the mean average weight of bears.

1. Random Quantitative Data? Yes. *This data was random and weight in pounds is a quantitative variable.*
2. Data values within the sample should be independent of each other. *This can be difficult to determine. It should not be the same bear measured multiple times. These bears were probably tagged so they probably did not accidentally measure the same bear multiple times. Also, one bear's weight should not change the probability of other bears having a certain weight. This data may not pass this assumption. Let us assume we see a bear that is eating well and is very heavy. Then there may be a higher probability of other bears being heavy in the same area.*
3. The sample data must be nearly normal or the sample size must be at least 30? *We see from the histogram that this data was skewed right, but the sample size was 54 (at least 30). Therefore, it does pass the 30 or normal requirement. Remember only one of the two need to be true for it to pass.*

The data did satisfy the random requirement and the at least 30 or normal requirement. If the data does satisfy the independence assumption, then the data would satisfy the overall requirements for using the formula and so the confidence interval will be relatively accurate.

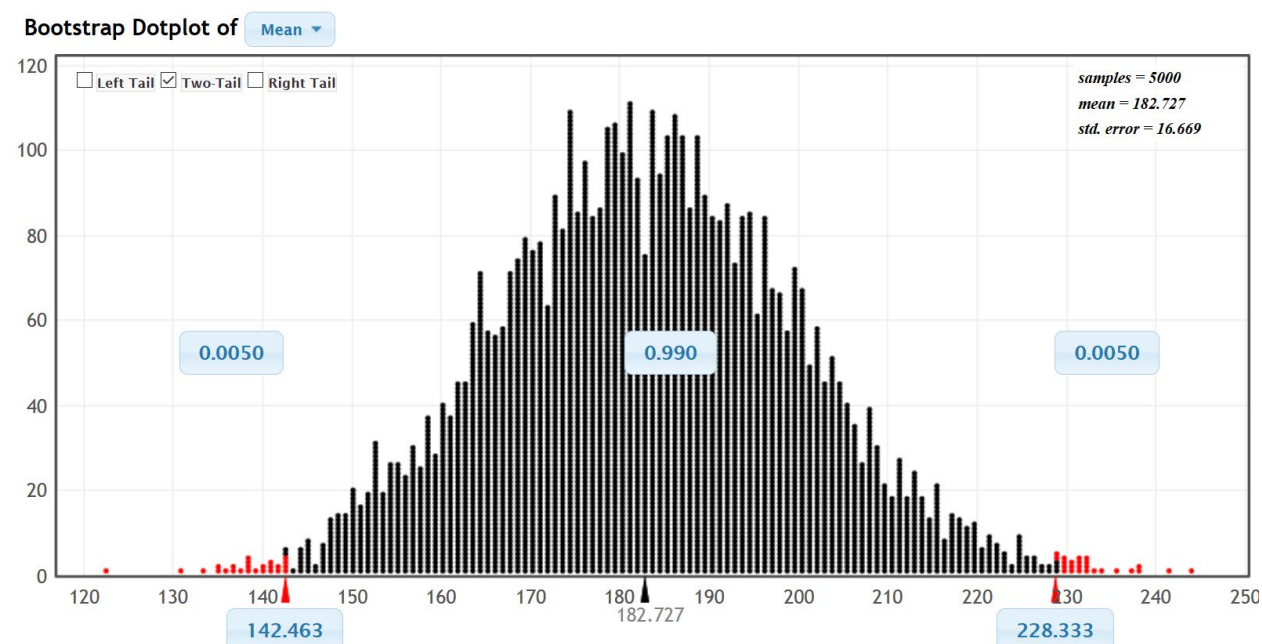
Could we have calculated this confidence interval with a bootstrap distribution?

Remember, the accuracy of a bootstrap is tied to the quality of the original sample data set. This data set was collected randomly but may fail the independence requirement.

Bear Weight Example

In this last example, we used the traditional T critical value and standard error formula to create a confidence interval and estimate the population mean average weight of bears. We could also use a bootstrap. First, go to the "Bear Data" at www.matt-teachout.org and copy the bear weight column of data. Now go to the "Bootstrap Confidence Interval" menu in StatKey at www.lock5stat.com and click on "CI for Single Mean, Median, St.Dev." Under "Edit Data", paste in the raw quantitative bear weight data. Make sure to check the "Header Row" box since this data set had a title and push "OK". Click "Generate 1000 Samples" a few times. Now click "Two-Tail".

The default is 95%, but you can change the middle proportion to 99%. This problem was a 99% confidence interval, so we will change the middle proportion to 99% (0.99).



We see that the bootstrap distribution is normally distributed. The confidence interval has a lower limit of 142.463 pounds, an upper limit of 228.333 pounds and a standard error of 16.669. Notice these numbers are relatively close to the same numbers we got by formula and Statcato. Since the confidence interval is normal, we can use the margin of error back-solving formula to find the approximate margin of error.

$$\text{Margin of Error} = \frac{(\text{Upper Limit} - \text{Lower Limit})}{2} = \frac{(228.333 - 142.463)}{2} \approx 42.935$$

Problem Set Section 2E

Directions: Answer the following questions.

1. What are the assumptions necessary for making a one-population proportion confidence interval?
2. What are the assumptions necessary for making a one-population mean confidence interval?
3. What are the assumptions necessary for making a one-population bootstrap confidence interval?
4. An experiment was conducted to see what percentage of rats would show empathy toward fellow rats in distress. Of the 30 total rats in the study, 23 showed empathy. What was the sample proportion? What are the critical value Z-scores for 99% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to “theoretical distributions” and click on “normal”. You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 99% confidence interval estimate of the population proportion of rats that show empathy. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.07725

- a) Sample Proportion $\hat{p} = \frac{\text{Number of Success (events)}}{\text{Total Sample Size}} =$
- b) Critical value Z-scores = \pm
- c) Margin of Error = $Z \times \text{Standard Error} =$
- d) Confidence Interval Lower Limit = $\hat{p} - (\text{Margin of Error})$
- e) Confidence Interval Upper Limit = $\hat{p} + (\text{Margin of Error})$

5. Use the following Statcato printout to check your margin of error and confidence interval answers from the rat empathy data in number 4. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	99.0%CI
30	23	0.767	0.199	(0.5678, 0.9656)

- a) Check each of the assumptions for this problem. Assume the rats were randomly selected. Explain your answers.
- b) Write a sentence to explain the margin of error in context.
- c) Write a sentence to explain the confidence interval in context.



6. A study was done on the effectiveness of lie detector tests to catch someone that lies. In a random sample of 48 total lies, the machine identified only 31 of them. What was the sample proportion? What are the critical value Z-scores for 95% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "normal". You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 95% confidence interval estimate of the population proportion of lies caught by lie detector tests. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.0689

- Sample Proportion $\hat{p} = \frac{\text{Number of Success (events)}}{\text{Total Sample Size}} =$
- Critical value Z-scores = \pm
- Margin of Error = $Z \times \text{Standard Error} =$
- Confidence Interval Lower Limit = $\hat{p} - (\text{Margin of Error})$
- Confidence Interval Upper Limit = $\hat{p} + (\text{Margin of Error})$

7. Use the following Statcato printout to check your margin of error and confidence interval answers from the lie detector data in number 6. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	95.0%CI
48	31	0.646	0.135	(0.5105, 0.7811)

- Check each of the assumptions for this problem. Explain your answers.
- Write a sentence to explain the margin of error in context.
- Write a sentence to explain the confidence interval in context.

8. We want to determine what percentage of cereals the company Quaker makes. A random sample of 24 cereals found that Quaker made four of them. What was the sample proportion? What are the critical value Z-scores for 90% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to "theoretical distributions" and click on "normal". You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 90% confidence interval estimate of the population proportion of cereals made by Quaker. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.076

- Sample Proportion $\hat{p} = \frac{\text{Number of Success (events)}}{\text{Total Sample Size}} =$
- Critical value Z-scores = \pm
- Margin of Error = $Z \times \text{Standard Error} =$
- Confidence Interval Lower Limit = $\hat{p} - (\text{Margin of Error})$
- Confidence Interval Upper Limit = $\hat{p} + (\text{Margin of Error})$



9. Use the following Statcato printout to check your margin of error and confidence interval answers from the cereal data in number 8. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	90.0%CI
24	4	0.167	0.125	(0.0415, 0.2918)

a) Check each of the assumptions for this problem. Assume the cereal data was randomly selected. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.

10. If a cereal has more than 9 grams of sugar per serving, we consider it to have a high sugar content. We want to determine what percentage of cereals have a high sugar content. A random sample of 24 cereals found that 10 of them have a high sugar content. What was the sample proportion? What are the critical value Z-scores for 95% confidence? If you cannot remember them, open StatKey at www.lock5stat.com. Go to "theoretical distributions" and then click on "normal". You can look up the critical value Z-scores. Use the critical values and the given standard error to calculate the margin of error and construct a 95% confidence interval estimate of the population proportion of cereals made by Quaker. Convert the upper and lower limits of your confidence interval into percentages.

Standard Error ≈ 0.1006

a) Sample Proportion $\hat{p} = \frac{\text{Number of Success (events)}}{\text{Total Sample Size}} =$

b) Critical value Z-scores = \pm

c) Margin of Error = $Z \times \text{Standard Error} =$

d) Confidence Interval Lower Limit = $\hat{p} - (\text{Margin of Error})$

e) Confidence Interval Upper Limit = $\hat{p} + (\text{Margin of Error})$

11. Use the following Statcato printout to check your margin of error and confidence interval answers from the cereal data in number 10. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Number of trials	Number of Events	Sample proportion	Margin of Error	95.0%CI
24	10	0.417	0.197	(0.2194, 0.6139)

a) Check each of the assumptions for this problem. Assume the cereal data was randomly selected. Explain your answers.

b) Write a sentence to explain the margin of error in context.

c) Write a sentence to explain the confidence interval in context.



12. A random sample of 45 high school students has a skewed left distribution. The sample mean average ACT exam score (\bar{x}) was 20.8 with a sample standard deviation of 9.868. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to “theoretical distributions” and click on “T”. Use the degrees of freedom and StatKey to look up the critical value T-scores for a 90% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 90% confidence interval estimate of the population mean average ACT exam.

Standard Error = 1.471 ACT points

- a) Degrees of Freedom = $n - 1 =$
- b) Critical value T-scores = \pm
- c) Margin of Error = $T \times \text{Standard Error} =$
- d) Confidence Interval Lower Limit = $\bar{x} - (\text{Margin of Error})$
- e) Confidence Interval Upper Limit = $\bar{x} + (\text{Margin of Error})$

13. Use the following Statcato printout to check your margin of error and confidence interval answers from the ACT data in number 12. Now check the assumptions and write sentences to explain the margin of error and confidence interval.

Var	N	Mean	Stdev	Margin of Error	90.0%CI
summary	45.0	20.8	9.868	2.472	(18.3284, 23.2716)

- a) Check each of the assumptions for this problem. Explain your answers.
- b) Write a sentence to explain the margin of error in context.
- c) Write a sentence to explain the confidence interval in context.

14. A random sample of body temperatures in degrees Fahrenheit was taken from 50 randomly selected adults. The sample mean temperature of 98.26 °F and a standard deviation of 0.765 °F. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to “theoretical distributions” and click on “T”. Use the degrees of freedom and StatKey to look up the critical value T-scores for a 95% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 95% confidence interval estimate of the population mean average body temperature.

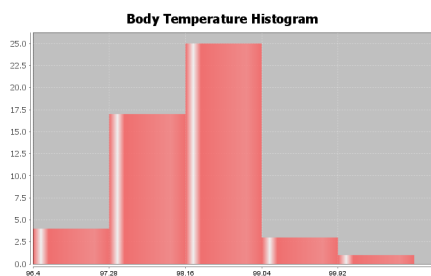
Standard Error = 0.1082 °F

- a) Degrees of Freedom = $n - 1 =$
- b) Critical value T-scores = \pm
- c) Margin of Error = $T \times \text{Standard Error} =$
- d) Confidence Interval Lower Limit = $\bar{x} - (\text{Margin of Error})$
- e) Confidence Interval Upper Limit = $\bar{x} + (\text{Margin of Error})$



15. Use the following Statcato printout to check your margin of error and confidence interval answers from the temperature data in number 14. Now check the assumptions and write sentences to explain the margin of error and confidence interval. A histogram of the data has been created with Statcato.

Var	N	Mean	Stdev	Margin of Error	95.0%CI
summary	50.0	98.26	0.765	0.217	(98.0426, 98.4774)



- Check each of the assumptions for this problem. Explain your answers.
- Write a sentence to explain the margin of error in context.
- Write a sentence to explain the confidence interval in context.

16. A random sample of cereal sugar content (grams per serving) was taken from 24 cereals. The sample mean average amount of sugar was of 7.208 grams per serving and a standard deviation of 4.634 grams per serving. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to “theoretical distributions” and click on “T”. Use the degrees of freedom and StatKey to look up the critical value T-scores for a 99% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 99% confidence interval estimate of the population mean average amount of sugar in cereals.

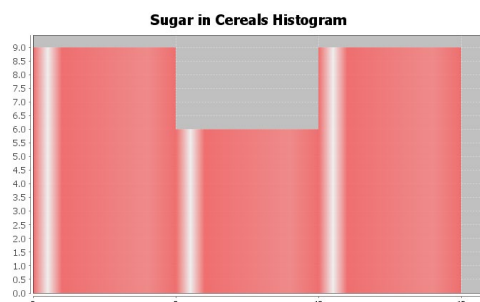
Standard Error = 0.9459 grams

- Degrees of Freedom = $n - 1 =$
- Critical value T-scores = \pm
- Margin of Error = $T \times \text{Standard Error} =$
- Confidence Interval Lower Limit = $\bar{x} - (\text{Margin of Error})$
- Confidence Interval Upper Limit = $\bar{x} + (\text{Margin of Error})$

17. Use the following Statcato printout to check your margin of error and confidence interval answers from the sugar in cereals data in number 16. Now check the assumptions and write sentences to explain the margin of error and confidence interval. A histogram of the data has been created with Statcato.

Var	N	Mean	Stdev	Margin of Error	99.0%CI
Sugar (grams per serving)	24.0	7.208	4.634	2.656	(4.5527, 9.8639)





- a) Check each of the assumptions for this problem. Explain your answers.
- b) Write a sentence to explain the margin of error in context.
- c) Write a sentence to explain the confidence interval in context.

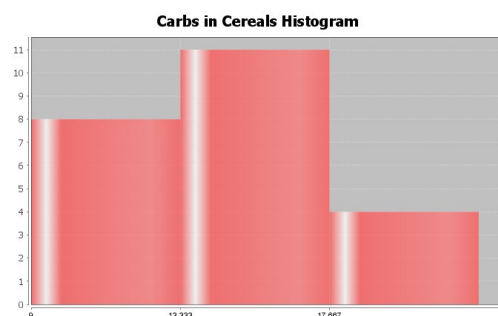
18. A random sample of cereal carbohydrate content (grams per serving) was taken from 24 cereals. The sample mean average amount of carbs was of 15.043 grams per serving and a standard deviation of 3.596 grams per serving. What is the degrees of freedom? Open StatKey at www.lock5stat.com. Go to “theoretical distributions” and click on “T”. Use the degrees of freedom and StatKey to look up the critical value T-scores for a 99% confidence level. Use the critical values and the given standard error to calculate the margin of error and construct a 99% confidence interval estimate of the population mean average amount of sugar in cereals.

Standard Error = 0.734 grams

- a) Degrees of Freedom = $n - 1 =$
- b) Critical value T-scores = \pm
- c) Margin of Error = $T \times \text{Standard Error} =$
- d) Confidence Interval Lower Limit = $\bar{x} - (\text{Margin of Error})$
- e) Confidence Interval Upper Limit = $\bar{x} + (\text{Margin of Error})$

19. Use the following Statcato printout to check your margin of error and confidence interval answers from the carbohydrates in cereals data in number 18. Now check the assumptions and write sentences to explain the margin of error and confidence interval. A histogram of the data has been created with Statcato.

Var	N	Mean	Stdev	Margin of Error	99.0%CI
summary	24.0	15.043	3.596	2.061	(12.9824, 17.1036)



- a) Check each of the assumptions for this problem. Explain your answers.
- b) Write a sentence to explain the margin of error in context.
- c) Write a sentence to explain the confidence interval in context.

One-Population Bootstrap Confidence Interval Practice Problems

20. An experiment was conducted to see what percentage of rats would show empathy toward fellow rats in distress. Of the 30 total rats in the study, 23 showed empathy. Go to www.lock5stat.com and click on the “StatKey” tab. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Single Proportion”. Click on “Edit Data” and enter 23 for the “count” and 30 for the “sample size”. Click on “Generate 1000 Samples” a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 99% confidence interval for the population proportion.

- a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.
- b) How many bootstrap samples did you take?
- c) What is the shape of the bootstrap distribution?
- d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #5. Are the close?
- e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

21. A study was done on the effectiveness of lie detector tests to catch someone that lies. In a random sample of 48 total lies, the machine identified only 31 of them. Go to www.lock5stat.com and click on the “StatKey” tab. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Single Proportion”. Click on “Edit Data” and enter 31 for the “count” and 48 for the “sample size”. Click on “Generate 1000 Samples” a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 95% confidence interval for the population proportion.

- a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.
- b) How many bootstrap samples did you take?
- c) What is the shape of the bootstrap distribution?
- d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #7. Are the close?
- e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

22. We want to determine what percentage of cereals the company Quaker makes. A random sample of 24 cereals found that Quaker made four of them. Go to www.lock5stat.com and click on the “StatKey” tab. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Single Proportion”. Click on “Edit Data” and enter 4 for the “count” and 24 for the “sample size”. Click on “Generate 1000 Samples” a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 90% confidence interval for the population proportion.

- a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.
- b) How many bootstrap samples did you take?
- c) What is the shape of the bootstrap distribution?
- d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #9. Are the close?



e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

23. If a cereal has more than 9 grams of sugar per serving, we consider it to have a high sugar content. We want to determine what percentage of cereals have a high sugar content. A random sample of 24 cereals found that 10 of them have a high sugar content. Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Proportion". Click on "Edit Data" and enter 10 for the "count" and 24 for the "sample size". Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the proportion. Use the bootstrap distribution to find a 95% confidence interval for the population proportion.

a) Does this data meet the assumptions for a bootstrap confidence interval? Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution?

d) Write the upper and lower limits of the bootstrap confidence interval. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #11. Are they close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population proportion.

24. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "sugar (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the sugar data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 99% confidence interval for the population mean.

a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.

b) How many bootstrap samples did you take?

c) What is the shape of the bootstrap distribution for the mean?

d) Write the upper and lower limits of the bootstrap confidence interval for the population mean. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #17. Are they close?

e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average amount of sugar in cereals. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 99% confidence interval for the population median.

f) What is the shape of the bootstrap distribution for the median?

g) Write the upper and lower limits of the bootstrap confidence interval for the population median.

h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.

25. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "carbs (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the carb data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 95% confidence interval for the population mean.



- a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.
- b) How many bootstrap samples did you take?
- c) What is the shape of the bootstrap distribution for the mean?
- d) Write the upper and lower limits of the bootstrap confidence interval for the population mean. Compare the upper and lower limits of the bootstrap confidence interval to the ones found by the traditional formula with Statcato in #19. Are the close?
- e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average amount of carbohydrates in cereals. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 95% confidence interval for the population median.

- f) What is the shape of the bootstrap distribution for the median?
- g) Write the upper and lower limits of the bootstrap confidence interval for the population median.
- h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.

26. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "bear data" in excel. Copy the column of data labeled "weight in pounds". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the bear weight data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 90% confidence interval for the population mean average weight of bears.

- a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.
- b) How many bootstrap samples did you take?
- c) What is the shape of the bootstrap distribution for the mean?
- d) Write the upper and lower limits of the bootstrap confidence interval for the population mean.
- e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average weight of bears. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 90% confidence interval for the population median.

- f) What is the shape of the bootstrap distribution for the median?
- g) Write the upper and lower limits of the bootstrap confidence interval for the population median.
- h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.

27. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "bear data" in excel. Copy the column of data labeled "length in inches". Do not click on "head length" by mistake. We want the overall length of the bears. Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Mean". Now click on "Edit Data" and paste the bear length data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the mean. Use the bootstrap distribution to find a 99% confidence interval for the population mean average length of bears.

- a) Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.



- b) How many bootstrap samples did you take?
- c) What is the shape of the bootstrap distribution for the mean?
- d) Write the upper and lower limits of the bootstrap confidence interval for the population mean.
- e) Write a sentence to explain the bootstrap confidence interval estimate of the population mean.

We can also use bootstrapping to estimate the population median average length of bears. Click on "Bootstrap Dot plot of Median". Use the bootstrap distribution to find a 99% confidence interval for the population median.

- f) What is the shape of the bootstrap distribution for the median?
 - g) Write the upper and lower limits of the bootstrap confidence interval for the population median.
 - h) Write a sentence to explain the bootstrap confidence interval estimate of the population median.
-

Section 2F – Two-Population Mean & Proportion Confidence Intervals

Studying the differences between two populations is very common in statistics, however sampling variability makes it very difficult to determine. Think of it this way. We know that random samples are usually different from each other, so even if two populations were the same, the samples taken from those populations would be different. A key question to ask is why are the samples different? Are the samples different because the populations are different or are my samples different because of sampling variability? Here is another key question. Are my samples significantly different or only slightly different? Two-population confidence intervals are often used to answer these difficult questions.

Before you can understand two-population confidence intervals, we have to take you back to arithmetic. A two-population confidence interval is the answer to a subtraction problem. Remember, the answer to a subtraction problem is often called the "difference". We need to understand how subtraction works and what a difference actually tells us.

Understanding Positive Differences

Suppose you subtract two numbers and the answer comes out positive. There is a positive difference. Is the first number bigger or smaller than the second number? Let us look at an example.

$$17 - 6 = +11$$

What does this tell us? Since the difference comes out positive, we know that the first number (17) is larger than the second number (6). It actually tells us more than this. The answer of +11 tells us that the first number (17) is 11 units larger than the second number (6).

How does this translate to a two-population confidence interval?

A two-population confidence interval does not measure population 1 or population 2 individually. Instead, it measures the difference between the population parameters. Two-population mean confidence intervals measure $\mu_1 - \mu_2$ (the difference between the population means). Two-population proportion confidence intervals measure $\pi_1 - \pi_2$ or $p_1 - p_2$ (the difference between the population proportions). The key is that the confidence interval is the answer to a subtraction problem.

Example: We want to compare the population mean height of men and women. We used a random sample of 40 men's heights in inches and a random sample of 40 women's heights in inches. Putting the data into Statcato, we got the following two population mean confidence interval. Population 1 was men's heights and population 2 was women's heights. We will assume for now that this data did meet the assumptions to estimate the populations.



Confidence Intervals - Two population means: confidence level = 0.95

Samples of population 1 in C16 Men Ht (in)

Samples of population 2 in C2 Women Ht (in)

	N	Mean	Stdev
Population 1	40	68.335	3.020
Population 2	40	63.195	2.741

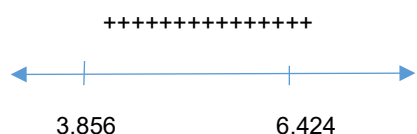
* Population standard deviations are unknown. *

DOF = 77

Margin of error = 1.284

95.0%CI = (3.8560, 6.4240)

First of all, 3.856 inches is NOT population 1 and 6.424 is NOT population 2. That is not how confidence intervals work. Remember this is an interval. It represents all of the numbers in between 3.856 and 6.424 inches and difference between the population means ($\mu_1 - \mu_2$) could be any of them. Think of the number line. Notice that all of the numbers between 3.856 and 6.424 are positive!



So while we do not know what the population difference is exactly, we do know that the difference is positive. Think back. Remember if the difference is positive, then the first number must be larger than the second. In this case, the mean average of population 1 (men's heights) is likely to be larger than the mean average of population 2 (women's heights). Remember the positive difference tells you how much larger.

Sentence to explain the confidence interval: We are 95% confident that the population mean average height of men (population 1) is between 3.856 and 6.424 inches larger than the population mean average height of women (population 2).

Note: You may also see the two-population confidence interval sentence written this way. We are 95% confident that the difference between the population mean average heights of men and women is between 3.856 and 6.424 inches. This can be a confusing way to explain the confidence interval though as people rarely understand the implications of that sentence.

Significance

Notice that the sample mean average height for the 40 men in the sample data was 68.335 inches and the sample mean average height for the 40 women in the sample data was 63.195 inches. Are these sample mean's significantly different? Yes. If both the upper and lower limits of your two population confidence interval are positive (+, +), then that does indicate that your sample statistic from group 1 is significantly higher than the sample statistic from group 2.

Positive Difference Two-population Confidence Intervals (+, +)

Sentence: "We are #% confident that the parameter from population 1 is between # and # larger than the parameter from population 2."



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Significance: There is a significant difference between the two samples. The sample statistic for group 1 is significantly higher than group 2. This indicates that the parameter for population 1 might be higher than for population 2.

Example 2: Suppose we want to compare the percentage of statistics students that are democrat and the percentage of statistics students that are republican. We used the fall 2015 COC survey data to create the following confidence interval. For now, we will assume the problem met the assumptions for estimating the populations. Population 1 was democratic COC statistics students and population 2 was republican COC statistics students. We used a 90% confidence level and Statcato to calculate the following two-population proportion confidence interval.

Confidence Interval - Two population proportions: confidence level = 0.9

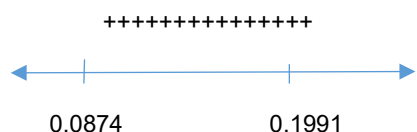
	Number of Events	Number of trials	Proportion
Sample 1	110	328	0.335
Sample 2	63	328	0.192

Sample proportion difference = 0.143

Margin of error = 0.056

90.0%CI = (0.0874, 0.1991)

Notice that both of the numbers in the two-population confidence interval are positive. These proportions can be converted to their percentage equivalent (+8.74% , +19.91%). Again 8.74% is NOT population 1 and 19.91% is NOT population 2. That is not how two-population confidence intervals work. The difference between the population proportions ($\pi_1 - \pi_2$) could be any of the numbers between 0.0874 and 0.1991. Notice again that all of the numbers in between 0.0874 and 0.1991 are positive.



So the population proportion difference $\pi_1 - \pi_2$ is positive. This tells us that the population proportion (and percentage) of COC statistics students that are democrat (population 1) is likely to be larger than the population proportion or percentage of COC statistics students that are republican (population 2). Remember the positive difference tells you how much larger.

Sentence:

We are 90% confident that the population percentage of COC statistics students that are democratic (population 1) is between 8.74% and 19.91% higher than the percentage of COC statistics students that are republican (population 2).

OR

We are 90% confident that the population proportion of COC statistics students that are democratic (population 1) is between 0.0874 and 0.1991 higher than the population proportion of COC statistics students that are republican (population 2).



Understanding Negative Differences

Suppose you subtract two numbers and the answer comes out negative. (There is a negative difference.) Is the first number bigger or smaller than the second number? Let us look at an example.

$$5 - 13 = -8$$

What does this tell us? Since the difference comes out negative, we know that the first number (5) is smaller than the second number (13). It actually tells us more than this. The answer of -8 tells us that the first number (5) is eight units smaller than the second number (13). Notice we did not say that the first number is -8 units smaller. The difference of -8 tells us that the first number is eight units smaller than the second number.

How does this translate to a two-population confidence interval?

Remember, a two-population confidence interval does not measure population 1 or population 2 individually. Instead, it measures the difference between the population parameters. Two-population mean confidence intervals measure $\mu_1 - \mu_2$ (the difference between the population means). Two-population proportion confidence intervals measure $\pi_1 - \pi_2$ or $p_1 - p_2$ (the difference between the population proportions).

Example: We want to compare the population mean weight of women and men. We used a random sample of 40 women's weights in pounds and a random sample of 40 men's weights in pounds. Putting the data into Statcato, we got the following two population mean confidence interval. Population 1 was women's weights and population 2 was men's weights. We will assume for now that this data did meet the assumptions to estimate the populations.

Confidence Intervals - Two population means: confidence level = 0.95

Samples of population 1 in C3 Women Wt (Lbs)

Samples of population 2 in C17 Men Wt (Lbs)

	N	Mean	Stdev
Population 1	40	146.220	37.621
Population 2	40	172.55	26.327

* Population standard deviations are unknown. *

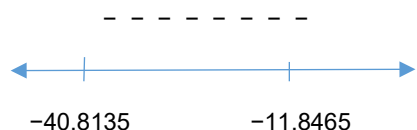
DOF = 69

Margin of error = 14.484

95.0%CI = (-40.8135, -11.8465)

Remember, -40.8135 pounds is NOT population 1 and -11.8465 pounds is NOT population 2. That is not how two-population confidence intervals work. Remember this is an interval. It represents all of the numbers in between -40.8135 and -11.8465 pounds and difference between the population means ($\mu_1 - \mu_2$) could be any of them. Notice that the lower limit is now -40.8135 on the left and the upper limit is -11.8465 on the right. Many students are confused by this, but that is how the number line works. The more negative a number is, the smaller it is. Therefore, -40.8135 is smaller -11.8465 .

How do we interpret this? Think again of the number line. Notice that all of the numbers between -40.8135 and -11.8465 are negative!



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/18

So while we do not know what the population difference is exactly, we do know that the difference is negative. Remember if the difference is negative, then the first number must be smaller than the second number. In this case, the mean average of population 1 (women's weights) is likely to be smaller than the mean average of population 2 (men's weights). The negative difference tells you how much smaller.

Sentence to explain the confidence interval: We are 95% confident that the population mean average weight of women (population 1) is between 11.8465 pounds and 40.8135 pounds less than the population mean average height of men (population 2).

Significance

Notice that the sample mean average weight for the 40 women in the sample data was 146.220 pounds and the sample mean average height for the 40 men in the sample data was 172.55 pounds. Are these sample mean's significantly different? Yes. If both the upper and lower limits of your two population confidence interval are negative $(-, -)$, then that does indicate that your sample statistic from group 1 is significantly lower than the sample statistic from group 2.

Negative Difference Two-population Confidence Intervals $(-, -)$

Sentence: "We are #% confident that the parameter from population 1 is between # and # lower than (or less than) the parameter from population 2."

Significance: There is a significant difference between the two samples. The sample statistic for group 1 is significantly lower than group 2. This indicates that parameter for population 1 is probably lower than for population 2.

Example 2: In a previous example, we compared the percentage of statistics students that are democrat and the percentage of statistics students that are republican. We assigned democrat to be population 1 and republican to be population 2. What would happen if we reverse that? Suppose we let population 1 to be republican COC statistics students and population 2 to be democrat COC statistics students. We used a 90% confidence level and Statcato to calculate the following two-population proportion confidence interval. Assume the problem met the assumptions for estimating the populations.

Confidence Interval - Two population proportions: confidence level = 0.9

	Number of Events	Number of trials	Proportion
Sample 1	63	328	0.192
Sample 2	110	328	0.335

Sample proportion difference = -0.143

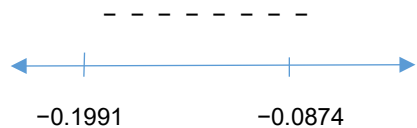
Margin of error = 0.056

90.0%CI = (-0.1991, -0.0874)

Notice that the sample difference is now negative, but the margin of error is the same. Both of the numbers in the two-population confidence interval are now negative. These proportions can be converted to their percentage equivalent $(-19.91\%, -8.74\%)$. Notice that these are the same percentages, but have opposite signs. Notice that the lower limit is now -19.91% on the left and the upper limit is -8.74% on the right. Remember, the more negative a



number is, the smaller it is, so -19.91% is smaller -8.74% . The difference between the population proportions ($\pi_1 - \pi_2$) could be any of the numbers between -0.1991 and -0.0874 . Notice again that all of the numbers in between -0.1991 and -0.0874 are negative.



So the population proportion difference $\pi_1 - \pi_2$ is negative. This tells us that the population proportion (and percentage) of COC statistics students that are republican (population 1) is likely to be smaller than the population proportion or percentage of COC statistics students that are democrat (population 2). The confidence interval being negative tells you how much smaller.

Sentence:

We are 90% confident that the population percentage of COC statistics students that are republican (population 1) is between 8.74% and 19.91% lower than the percentage of COC statistics students that are democrat (population 2).

OR

We are 90% confident that the population proportion of COC statistics students that are republican (population 1) is between 0.0874 and 0.1991 lower than the population proportion of COC statistics students that are democrat (population 2).

Significance:

Since both the upper and lower limits of the confidence interval were negative, this suggests that the sample percentage for group 1 (republican) was significantly lower than the sample percentage for group 2 (democrat). This indicates that the population percentage for republican COC statistics students is likely to be lower than the percentage for democratic COC statistics students.

Zero Difference

If we subtract two numbers and the answer is zero, the two numbers must be the same. Look at the following example.

$$13 - 13 = 0$$

The zero difference tells us that the first number (13) is the same as the second number (13).

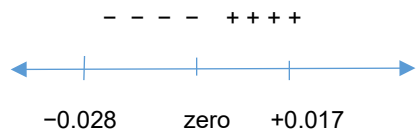
Example: Suppose 95% two-population proportion confidence interval came out to be $(-0.028, +0.017)$. Remember, -0.028 is NOT population 1 and $+0.017$ is NOT population 2. This confidence interval tells us that the difference between the population proportions ($\pi_1 - \pi_2$) is somewhere between -0.028 and $+0.017$. Some people will write the sentence as follows.

Sentence: We are 95% confident that the population proportion difference is between -0.028 and $+0.017$.

What does that even mean? Is population 1 lower or higher than population 2? How much lower or higher? To answer these questions, we need to examine the number line between -0.028 and $+0.017$. Notice that there are many negative numbers in this interval, so population 1 may be lower than population 2. There are also many positive numbers in this interval, so population 1 may be higher than population 2. Zero is also in the interval, so it is also a possibility. Remember if the difference is zero, then population 1 and population 2 could be the same. This interval tells us that we really do not know which population is larger. When the upper and lower limits of a two-



population confidence interval have opposite signs, this means there is no significant difference between the populations. The sample statistics for the two groups are so close, that we cannot tell if population 1 is lower or higher than population 2. They could be the same.

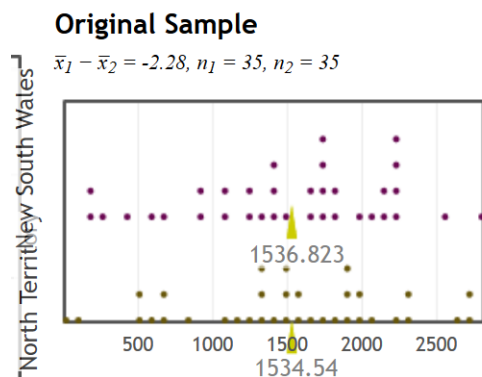


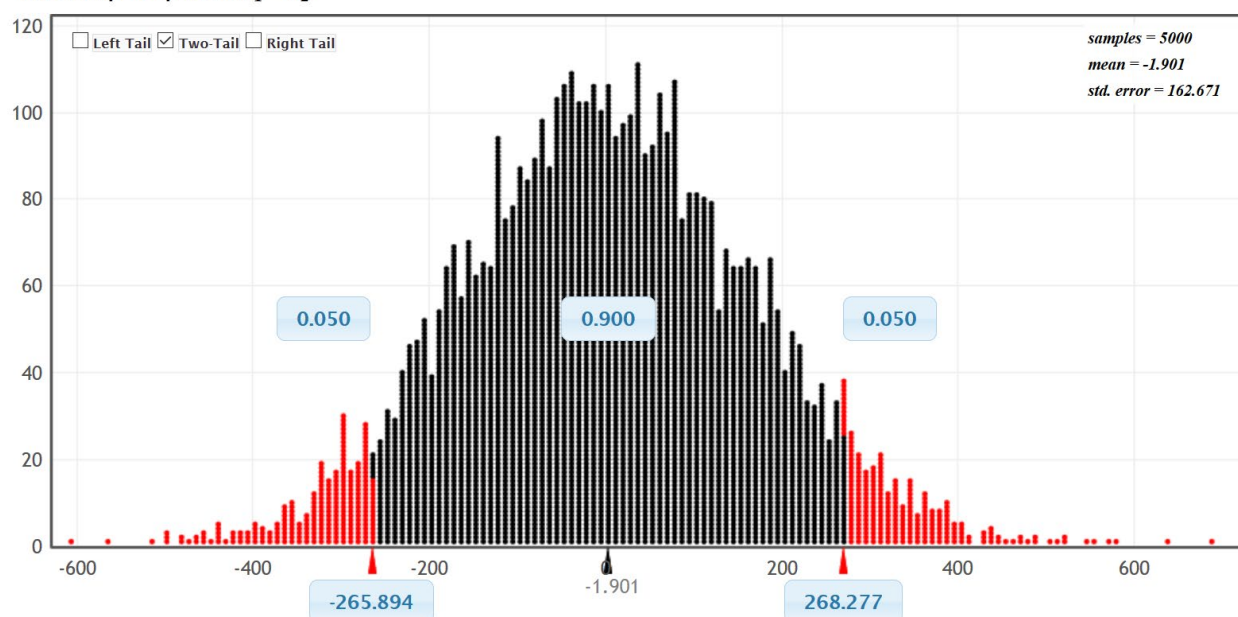
Two-population Confidence Intervals (- , +)

Sentence: We are #% confident that there is no significant difference between the parameter for population 1 and parameter for population 2.

Significance: When the upper and lower limits for the two-population confidence interval have opposite signs, then that indicates that the sample statistics for the two groups are not significantly different.

Example: Let us compare the population mean average salary of people living in Northern Territory, Australia (μ_1) to people living in New South Wales, Australia (μ_2). We used StatKey and random sample data to create the following two-population mean bootstrap 90% confidence interval. Assume the data met all of the assumptions.



Bootstrap Dotplot of $\bar{x}_1 - \bar{x}_2$ 

From the bootstrap, we see that the 90% confidence interval is $(-265.894, +268.277)$. Notice the upper and lower limit have opposite signs. This tells us that the sample mean average salary for people living in the Northern Territory (\$1534.54) is not significantly different from the sample mean average salary for people living in New South Wales (\$1536.82). Since our sample means are so close, we cannot tell which population has a higher population mean average salary.

Confidence Interval Sentence: We are 90% confident that the difference between the population mean average salary of people living in Northern Territory, Australia and those living in New South Wales, Australia is between $-\$265.894$ and $+\$268.277$. (*This sentence tends to be confusing.*)

Better Confidence Interval Sentence: We are 90% confident that there is no significant difference between the population mean average salary of people living in Northern Territory, Australia and those living in New South Wales, Australia.

Note: We could also have calculated the 90% confidence interval with Statcato. Notice the upper and lower limits of the bootstrap are similar to what Statcato calculated.



Confidence Intervals - Two population means: confidence level = 0.9

Samples of population 1 in C1 North Territory ...

Samples of population 2 in C2 New South Wales ...

	N	Mean	Stdev
Population 1	35	1534.540	701.525
Population 2	35	1536.823	677.140

* Population standard deviations are unknown. *

DOF = 67

Margin of error = 274.883

90.0%CI = (-277.1660, 272.5998)

Calculating Two-population Mean and Proportion Confidence Intervals

We will now discuss the formulas and calculations for two-population mean and proportion confidence intervals. It is important to understand the formulas and be able to explain them. However, no statistician or data scientist calculates these by hand with a formula. We virtually always use computer software to calculate any difficult calculations like confidence intervals.

Two-population Mean Confidence Intervals

There are two types of two-population mean confidence intervals, independent groups and matched pairs. Matched pair data is a one-to-one pairing between the two groups. Matched pair data usually from the same person measured twice. For example, the first number in the first data set comes from the same person as the first number in the second data set. The second numbers in each data set come from the same person and so on. Matched pairs do not have to be the same person measured twice. It could also be comparing husbands and wives, or sisters and brothers. You could be comparing two football teams and comparing the salary for each position: the starting quarterbacks, the starting running backs, the starting right guard, etc. Notice that in matched pairs, the sample sizes for the two groups are the same.

Use independent groups when you are comparing separate groups. For example, like comparing a random sample of men to a random sample of women or comparing a random sample of people from California to a random sample of people from Arizona.

Example 1 (Matched Pair): Let's use the random sample health data and a 99% confidence interval to compare the population mean systolic and diastolic blood pressure for men. Since these values come from the same 40 men, they are matched pairs.

Population 1: Men's Systolic Blood Pressure (mm of Hg)

Population 2: Men's Diastolic Blood Pressure (mm of Hg)

For independent groups, we calculate sample mean and sample standard deviation separately for each group and then subtract the sample means. For matched pair, we subtract the ordered pairs first, and then calculate the mean of the difference (\bar{d}) and the standard deviation of the difference (s_d).



Since the systolic and diastolic blood pressure for these 40 men were matched pairs, notice we subtracted each pair and created a new column of data called the “difference” column. A two-population mean matched pair confidence interval is calculating a one-population confidence interval using just the difference column.

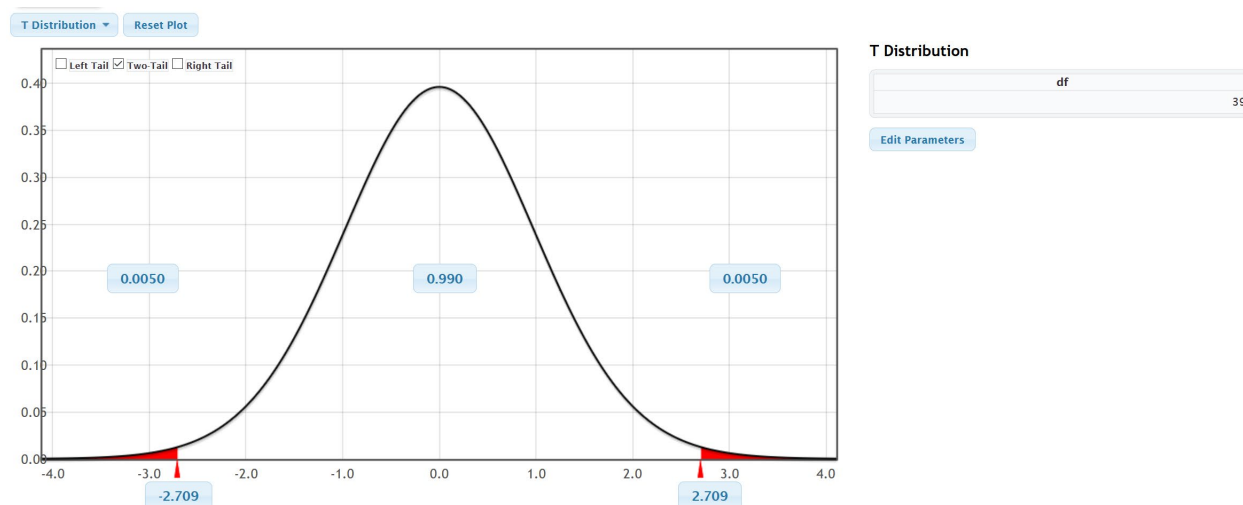
Men Syst BP (mm of Hg)	Men Diast BP (mm of Hg)	Difference between pairs
125	78	47
107	54	53
126	81	45
110	68	42
110	66	44
107	83	24
113	71	42
126	72	54
137	85	52
110	71	39
109	65	44
153	87	66
112	77	35
119	81	38
113	82	31
125	76	49
131	80	51
121	75	46
132	81	51
112	44	68
121	65	56
116	64	52
95	58	37
110	70	40
110	66	44

Descriptive Statistics

Variable	Mean	Standard Deviation
C3 Difference between pairs	45.675	9.352

The sample mean of the difference (\bar{d}) is 45.675, the sample standard deviation of the difference (s_d) is 9.352 and the sample size (n) is 40. These are used in the confidence interval calculation. We will also need to look up the T critical value. In matched pair, the sample size is the number of pairs (40), so the degrees of freedom is 39. We can use the theoretical T distribution function in StatKey to look up the critical value.





Notice that the T critical values are ± 2.709 . Here is the formula and calculation for the two-population mean matched pair confidence interval. Notice that the sample mean difference is 45.675 mm of Hg and the margin of error is 4.0057 mm of Hg. This gave us a confidence interval of

$$\bar{d} \pm T \frac{s_d}{\sqrt{n}}$$

$$45.675 \pm 2.709 \frac{9.352}{\sqrt{40}}$$

$$45.675 \pm 4.0057$$

$$(41.6693, 49.6807)$$

Notice that the upper and lower limits of the confidence interval are both positive. This tells us that the population 1 (men's systolic blood pressure) is higher than population 2 (men's diastolic blood pressure).

Sentence: We are 99% confident that the population mean systolic blood pressure for men is between 41.67 mm of Hg and 49.68 mm of Hg higher than the population mean diastolic blood pressure for men.

We can use Statcato to calculate this for us. Just go to the "statistics" menu in Statcato, click on "confidence intervals" and then matched pair. You can put in the summary data (sample mean difference 45.675, sample standard deviation of the differences 9.352, and sample size 40). You can also copy and paste the two quantitative data sets and then click the "samples in two columns" button.

Statcato => Statistics => Confidence Intervals => Matched Pairs

Confidence Interval - Matched Pairs: confidence level = 0.99

Sample 1: C1 Men Syst BP (mm ...

Sample 2: C2 Men Diast BP (mm...

Difference of Matched Pairs C1 Men Syst BP (mm ... - C2 Men Diast BP (mm...

N	Mean	Stdev	Margin of Error	99.0%CI
40	45.675	9.352	4.004	(41.6710, 49.6790)

Notice that the confidence interval in Statcato is virtually the same as our formula calculation above.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

We can also use bootstrapping in StatKey to calculate this confidence interval. Remember a matched pair is calculated as a one-population mean bootstrap from the differences between the pairs. Let us start by calculating the difference column in Excel. Copy and pasted the two data sets into excel. In cell "C2" type in " $=B2-C2$ " and push enter. Hold your curser on the bottom right corner until it turns into a "+". Double click and the formula will be applied to the rest of the data. You can also click and drag.



C2				=A2-B2	
	A	B	C		
1	Men Syst BP (mm of Hg)	Men Diast BP (mm of Hg)	Difference (Systolic - Diastolic)		
2	125	78	47		
3	107	54	53		
4	126	81	45		
5	110	68	42		
6	110	66	44		
7	107	83	24		
8	113	71	42		
9	126	72	54		
10	137	85	52		
11	110	71	39		
12	109	65	44		
13	153	87	66		
14	112	77	35		
15	119	81	38		
16	113	82	31		
17	125	76	49		
18	131	80	51		
19	121	75	46		
20	132	81	51		
21	112	44	68		
22	121	65	56		
23	116	64	52		
24	95	58	37		
25	110	70	40		
26	110	66	44		
27	125	82	43		
28	124	79	45		
29	131	69	62		
30	109	64	45		
31	112	79	33		
32	127	72	55		
33	132	74	58		
34	116	81	35		
35	125	84	41		
36	112	77	35		
37	125	77	48		
38	120	83	37		
39	118	68	50		
40	115	75	40		
41	115	65	50		
42					

Open StatKey at www.lock5stat.com and click on “CI for Single Mean, Median, St.Dev.” under the “bootstrap confidence interval” menu. Make sure the bootstrap dot plot is set to “mean”. Click on edit data. Copy and paste the “difference” column only and push “OK”.



Edit data
✕

Difference (Systolic - Diastolic)

47
53
45
42
44
24
42
54
52
39
44
66
35
38
31
49
51
46
51
68
50

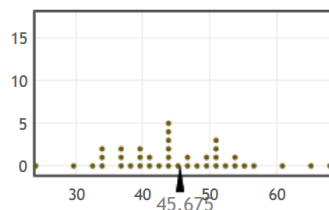
☐ First column is identifier
☒ Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only one column (or two if there is an identifier).

Ok

Original Sample

$n = 40$, mean = 45.675
median = 45, stdev = 9.352

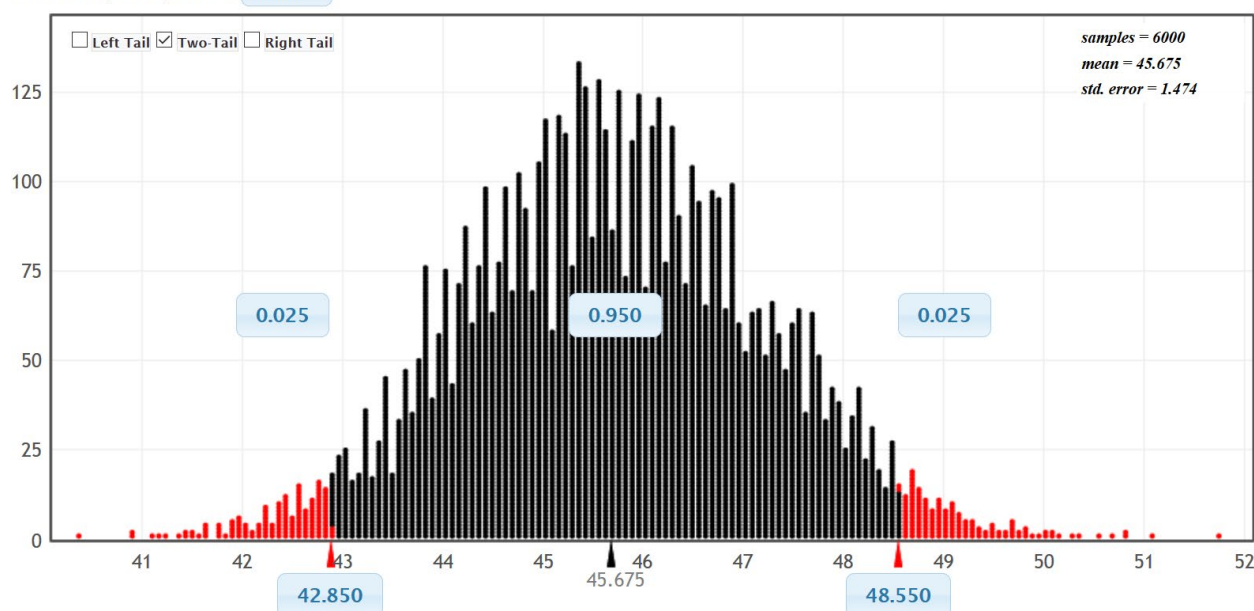


Now create the bootstrap distribution by clicking on “generate 1000 samples” a few times and click on two-tail. The default is 95% but you can change the middle proportion to 0.90 or 0.99 if needed. Notice the 95% bootstrap confidence interval is (+42.85 , +48.55). This is similar to our formula calculations above.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Bootstrap Dotplot of Mean



Important Notes about two-population bootstraps: Remember bootstrap confidence intervals will always come out slightly different because of sampling variability. Also that though we used a one-population bootstrap, this was not a one-population confidence interval. It measured the difference between the populations and must be interpreted accordingly. Remember to keep track of population 1 and population 2 and the signs of the confidence intervals.

Example 2 (Two-population mean from Independent Groups): Earlier we used the health data to calculate the following two-population confidence interval to compare the population mean average weight of women and men. Notice these groups are independent and not a one-to-one pairing. The upper and lower limits were negative, indicating that we are 95% confident that the population mean average weight of women is between 11.8465 pounds and 40.8135 pounds less than the population mean average weight of men.

Confidence Intervals - Two population means: confidence level = 0.95

Samples of population 1 in C3 Women Wt (Lbs)

Samples of population 2 in C17 Men Wt (Lbs)

	N	Mean	Stdev
Population 1	40	146.220	37.621
Population 2	40	172.55	26.327

* Population standard deviations are unknown. *

DOF = 69

Margin of error = 14.484

95.0%CI = (-40.8135, -11.8465)



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Let us discuss how Statcato calculated this confidence interval. Let us start with the degrees of freedom. For independent groups, the degrees of freedom calculation is much more difficult. There are many free online calculators for degrees of freedom. I like to use this one. You will need to enter the sample size and sample standard deviation for each of your two samples. Notice the degrees of freedom calculator gave 69.809. It is usually common to round down the degrees of freedom to account for possible greater variability. Notice Statcato rounded this degree of freedom down to 69 even though it was closer to 70.

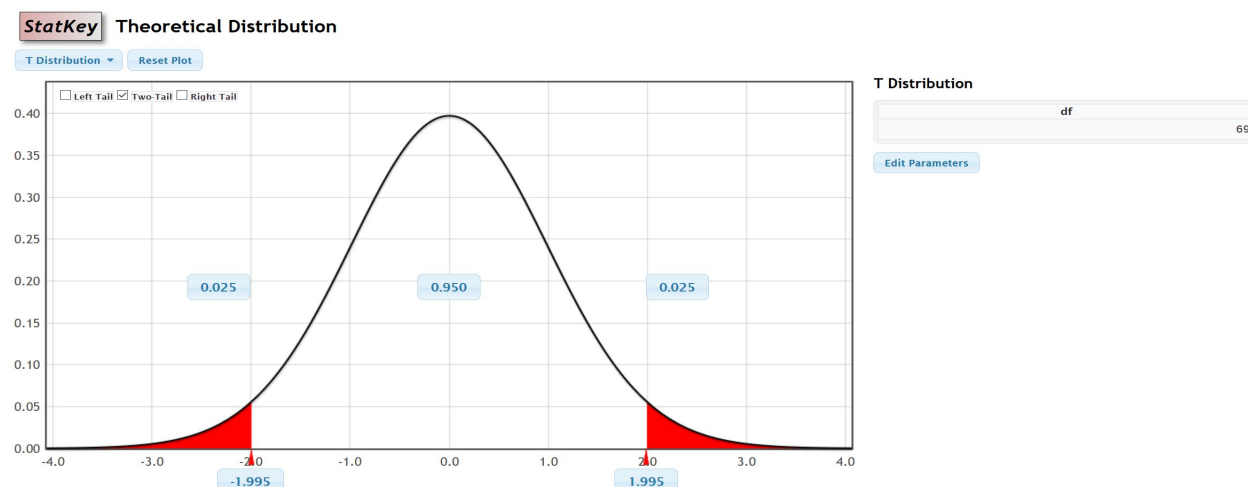
<http://web.utk.edu/~cwiek/TwoSampleDoF>

Compute Degrees of Freedom for t-test comparing means of two independent samples

Enter in the sample sizes (n_1 , n_2) and sample standard deviations (s_1 , s_2) and click "Compute DF" to get the degrees of freedom describing the sampling distribution of the difference in sample means.

n_1 : 40 n_2 : 40 s_1 : 37.621 s_2 : 26.327 Compute DF DF: 69.809

We can now look up the critical value T-scores for this confidence interval with the Theoretical Distribution T-score calculator in StatKey. Notice the critical value T-scores for 69 degrees of freedom are ± 1.995 .



Here is the formula for the two-population mean confidence interval for independent groups. We see that the sample mean weight for the women (\bar{x}_1) was 146.220 pounds, the sample mean weight for the men (\bar{x}_2) was 172.55 pounds, the sample standard deviation for the women's weights (s_1) was 37.621 pounds and the sample standard deviation for the men's weights (s_2) is 26.327 pounds. While both sample sizes are 40 in this example, it is common for independent groups to have different sample sizes.

$$(\bar{x}_1 - \bar{x}_2) \pm T \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$(146.22 - 172.55) \pm 1.995 \sqrt{\left(\frac{37.621^2}{40} + \frac{26.327^2}{40}\right)}$$

$$-26.33 \pm 1.995 (7.26)$$

$$-26.33 \pm 14.48$$

$$(-40.81, -11.85)$$



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Note about Pooling the Variances: Statisticians sometimes pool the variances when comparing the population means from two populations. It requires the population variances to be the same. For students new to stats, it is better not to pool the variances.

We can also calculate this confidence interval with bootstrapping. Go to www.lock5stat.com and click on StatKey. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Difference in Means”. This menu is for independent groups. While Statcato prefers the data to be separated by group, StatKey prefers to have the categorical and quantitative data. Copy and paste the raw gender and weight data into a new excel spreadsheet first. They need to be next to each other. Now click on “Edit Data”. Copy and paste the two columns into StatKey and push “Ok”.

A	B
Gender	Weight (Lbs)
Female	114.8
Female	149.3
Female	107.8
Female	160.1
Female	127.1
Female	123.1
Female	111.7
Female	156.3
Female	218.8
Female	110.2
Female	188.3
Female	105.4
Female	136.1
Female	182.4
Female	238.4
Female	108.8
Female	119
Female	161.9
Female	174.1
Female	181.2
Female	124.3
Female	255.9
Female	106.7
Female	149.9
Female	163.1
Female	94.3
Female	159.7
Female	162.8
Female	130
Female	179.9
Female	147.8
Female	112.9
Female	195.6
Female	124.2
Female	135
Female	141.4
Female	123.9
Female	135.5
Female	130.4
Female	100.7
male	169.1
male	144.2
male	179.3
male	175.8
male	152.6
male	166.8
male	135
male	201.5
male	175.2
male	139
male	156.3
male	186.6
male	191.1
male	151.3
male	209.4
male	237.1
male	176.7
male	220.6
male	166.1
male	137.4
male	164.2
male	162.4
male	151.8
male	144.1
male	204.6
male	193.8
male	172.9
male	161.9
male	174.8
male	169.8



Edit data

Gender	Weight (Lbs)
Female	114.8
Female	149.3
Female	107.8
Female	160.1
Female	127.1
Female	123.1
Female	111.7
Female	156.3
Female	218.8
Female	110.2
Female	188.3
Female	105.4
Female	136.1
Female	182.4
Female	238.4
Female	108.8
Female	119
Female	161.9
Female	174.1
Female	181.2

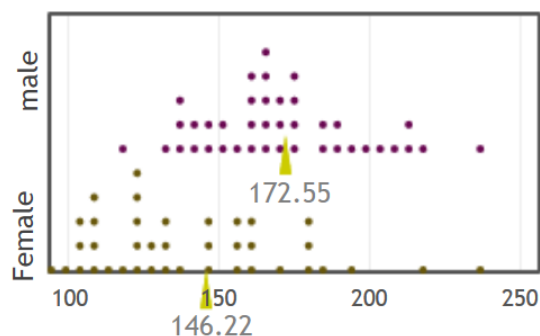
☒ Data has header row

Manually edit the values above or paste a tab or comma seperated file into the box and click Ok. The file must have only two columns where the first column is the categorical variable and the second is the quantitative.

Ok

Original Sample

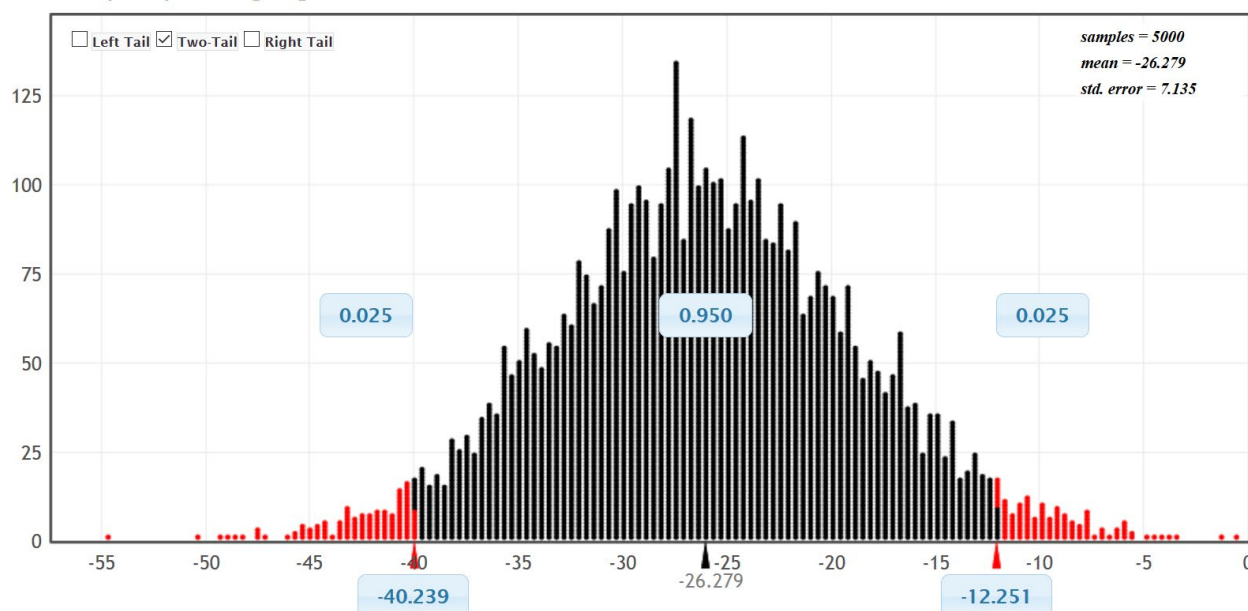
$$\bar{x}_1 - \bar{x}_2 = -26.33, n_1 = 40, n_2 = 40$$



Now create the bootstrap distribution by clicking on “generate 1000 samples” a few times and click on two-tail. The default is 95% but you can change the middle proportion to 0.90 or 0.99 if needed. Notice the 95% bootstrap confidence interval is $(-40.239, -12.251)$. This is similar to our formula calculations above.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Bootstrap Dotplot of $\bar{x}_1 - \bar{x}_2$ 

Important notes about two-population bootstraps: Remember bootstrap confidence intervals will always come out slightly different because of sampling variability. The two numbers at the bottom of the bootstrap distribution are the upper and lower limits of the confidence interval. For two-population, we need to keep track of population 1 and population 2 and the signs of the confidence intervals. In this case, population 1 was women's weights and population 2 was men's weights and the upper and lower limits were both negative.

Example 3 (Two-population proportion): Let's use the Math 140 survey data fall 2015 to compare the population percentage (proportion) of COC statistics students born in June (population 1) and the percentage (proportion) of COC statistics students born in December (population 2). We will assume the data met all of the assumptions for a two-population proportion confidence interval. The sample data showed that of the 336 total COC statistics students, 15 were born in June and 41 were born in December. We can use Statcato and a 90% confidence level to calculate the two-population proportion confidence interval. Just go to the "statistics" menu, and then click on "confidence intervals" and then "two-population proportion". Some computer programs will ask if you want to pool the samples. This means that you combine the two samples before calculating the standard error. Pooling is a technique used in hypothesis testing, but we do not pool the samples for two-population proportion confidence intervals.

Statcato => Statistics => Confidence Intervals => Two population Proportion

Note: Do not pool the sample proportions for confidence intervals.



Confidence Interval - Two population proportions: confidence level = 0.9

	Number of Events	Number of trials	Proportion
Sample 1	15	336	0.045
Sample 2	41	336	0.122

Sample proportion difference = -0.077

Margin of error = 0.035

90.0%CI = (-0.1121, -0.0427)

We see that the upper and lower limits of the confidence interval are both negative. This indicates that the proportion of COC statistics students born in June (population 1) is between 0.0427 and 0.1121 lower than the proportion of COC statistics students born in December (population 2).

Two-population proportion formula: Let us look at how this was calculated. Here is the two-population proportion formula. The sample proportion for group 1 (\hat{p}_1) was $15 \div 336 \approx 0.04464$ and the sample proportion for group 2 (\hat{p}_2) was $41 \div 336 \approx 0.12202$. Remember the famous Z-score critical value for 90% confidence is $Z = \pm 1.645$

$$(\hat{p}_1 - \hat{p}_2) \pm Z \sqrt{\left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)}$$

$$(0.04464 - 0.12202) \pm 1.645 \sqrt{\frac{0.04464(1 - 0.04464)}{336} + \frac{0.12202(1 - 0.12202)}{336}}$$

$$-0.07738 \pm 0.034731$$

$$(-0.1121, -0.04265)$$

Bootstrapping: We can also calculate this confidence interval with bootstrapping. Go to www.lock5stat.com and click on "StatKey". Under the "bootstrap confidence interval" section click on "difference in proportions". Click on the "Edit Data" button, and then enter the counts and sample sizes for both groups. Remember June was group 1 and December was group 2. Then push "Ok". Now generate a few thousand bootstrap samples, click two tail, and then put "0.90" for the middle proportion.

Edit data
✕

Please select values for two categories of count and sample size.

Group 1 count:

Group 1 sample size:

Group 2 count:

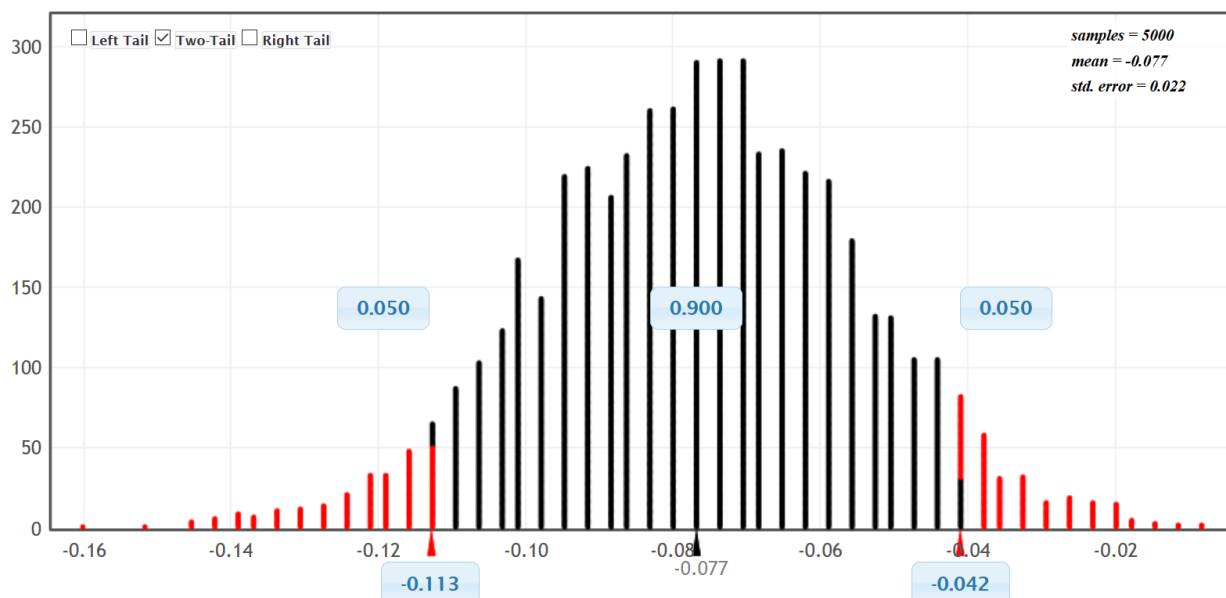
Group 2 sample size:



Original Sample

Group	Count	Sample Size	Proportion
Group 1	15	336	0.045
Group 2	41	336	0.122
Group 1-Group 2	-26	n/a	-0.077

Bootstrap Dotplot of $\hat{p}_1 - \hat{p}_2$



Notice the bootstrap distribution is normal and centered close to the sample proportion difference of -0.077 . It also indicates that the 90% confidence interval is $(-0.113, -0.042)$. This is close to what we got from the formula and Statcato.

Checking Assumptions

In order to compare populations, our sample data must be representative of the population. We usually require both samples to be large, random, and unbiased. The following assumptions are often used to check whether the sample data represents the population or not. Remember, if the sample data does not meet all of the assumptions, then we will not be able to draw any conclusions about the populations. It is also important to remember that these assumptions do not address all possible sources of bias.

Note about Independence:

- It is difficult to know for sure whether samples or individuals are indeed independent of each other. Individuals taken from two simple random samples from large populations will most likely be independent. A simple random sample of 50 people taken from a population of millions, will probably pass the individuals independent requirement. It is unlikely that we accidentally got people from the same family or people that work for the same company.



- Data collected conveniently or from voluntary response may fail the independence requirements. For example, if the sample data was collected from people in the same coffee shop or store, or on the same Facebook page, then they may be related or friends. This data would probably fail the independence requirements.
- Matched pair data means that there is a one-to-one pairing between the two samples. Usually it is the same people or objects measured twice. If the data is not matched pair, we often use the formulas for independent samples.

Notes about Bootstrapping:

- Bootstrapping does not require as many assumptions as traditional formula approaches and is often used when sample data fails the sample size requirements. However, bootstrapping does require the random and independence assumptions.

Notes about Experiments:

- Two-population confidence intervals can also be used in experimental design in order to prove cause and effect. In an experiment, the groups will not be random samples. They will need to be randomly assigned instead. Random assignment controls confounding variables.
- If the experiment uses random assignment, passes the assumptions, and shows a significant difference between the groups, then it indicates cause and effect.

Two-population Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape.

Two-population Mean Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

Two-population Proportion Assumptions

- The two categorical samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the samples should be independent of each other.
- There should be at least ten successes and at least ten failures.

Two-population Bootstrap Assumptions

- The sample data should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- If multiple samples were collected that were not matched pair, then the data values between the samples should be independent of each other.

Checking Assumptions Example 1: Earlier, we used the Math 140 survey data from fall 2015 to compare the population percentage (proportion) of COC statistics students born in June (population 1) and the percentage (proportion) of COC statistics students born in December (population 2). The sample data showed that of the 336 total COC statistics students, 15 were born in June and 41 were born in December. Would this data meet all of the assumptions for two-population confidence intervals with the traditional formula approach?



Traditional Formula Assumptions for Comparing Two-Population Proportions (Percentages)

- The two categorical samples should be collected randomly or be representative of the population.

No. The month a person is born in is categorical; however, the sample data was not collected randomly. It was a census of all statistics students in the fall 2015 semester. Occasionally, data that is not collected randomly may still be representative. If we consider our population of interest as all statistics students from all semesters then this data may still be representative of the population of interest, even though it is not random.

- Both samples should have at least 10 successes and at least 10 failures.

Yes. There were 15 students born in June and 41 in December. Both of these are greater than 10. There were $336 - 15 = 321$ students NOT born in June and $336 - 41 = 295$ NOT born in December. Both of these are greater than 10.

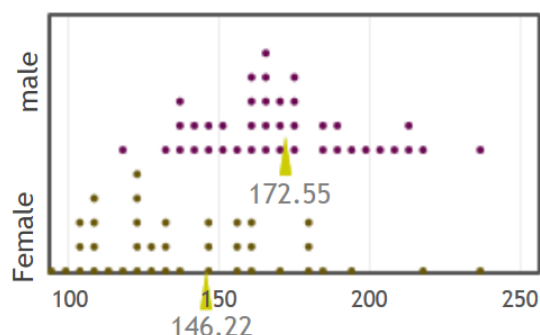
- Data values within each sample and between the samples should be independent of each other.

No. This data likely fails the independence requirements. Many students came from the same statistics classes.

Checking Assumptions Example 2: earlier in this section, we used the randomly collected health data to compare the population mean weight of women to the population mean weight of men. Were these groups matched pair? Would this data pass the traditional formula assumptions for comparing the means? The sample data is given below.

Original Sample

$$\bar{x}_1 - \bar{x}_2 = -26.33, n_1 = 40, n_2 = 40$$



A random sample of men and women are not matched pair. They were not husband and wife or brother and sister. Therefore, we will proceed with checking the assumptions for independent groups.

Traditional Formula Assumptions for Comparing Two-Population Means from Independent Groups

- Both samples should be random and quantitative.

Yes. Weights are quantitative and these were both random samples.

- Each sample should be Either Nearly Normal (almost bell shaped) OR have a Sample size of at least 30.



Women: Yes. The dot plot for the women's weight data shows that it is skewed right. So our sample size must be over 30 for it to pass. The sample size is 40, which is greater than the 30 requirement. Even though the shape was not normal, it still passes the at least 30 or normal requirement.

Men: Yes. The dot plot for the men's weight data shows that it is nearly normal. The sample size is 40, which is greater than the 30 requirement.

- Data values within the samples and between the samples should be independent of each other.

Yes. A small random sample of 40 men and 40 women taken from millions of men and women in the population, are not likely to be related or know each other.

Problems Section 2F

1. What are the assumptions we should check if we want to use sample data to calculate a two-population proportion confidence interval?
2. What are the assumptions we should check if we want to use sample data to calculate a two-population mean confidence interval?
3. What are the assumptions we should check if we want to use sample data to calculate a two-population bootstrap confidence interval?

(#4-12) Answer the following questions. Assume the confidence intervals met the assumptions.

4. A two-population mean confidence interval is (+3.4 kg , +5.9 kg). They used a 90% confidence level.
 - a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
 - b) How much higher could the population mean from population 1 than the population mean from population 2?
 - c) Write the two-population confidence interval sentence explaining this confidence interval.
5. A two-population proportion confidence interval is (−0.115 , −0.068). They used a 95% confidence level.
 - a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
 - b) How much lower could the percentage from population 1 be than the percentage from population 2?
 - c) Write the two-population confidence interval sentence explaining this confidence interval.
6. A two-population mean confidence interval is (−16.4°F , +8.2°F). They used a 99% confidence level.
 - a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
 - b) Write the two-population confidence interval sentence explaining this confidence interval.



7. A two-population proportion confidence interval is $(-0.045, +0.038)$. They used a 90% confidence level.
 - a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
 - b) Write the two-population confidence interval sentence explaining this confidence interval.
8. A two-population mean confidence interval is $(-\$185.71, -\$103.62)$. They used a 95% confidence level.
 - a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
 - b) How much lower could the population mean from population 1 than the population mean from population 2?
 - c) Write the two-population confidence interval sentence explaining this confidence interval.
9. A two-population proportion confidence interval is $(+0.049, +0.058)$. They used a 99% confidence level.
 - a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
 - b) How much higher could the percentage from population 1 be than the percentage from population 2?
 - c) Write the two-population confidence interval sentence explaining this confidence interval.
10. A two-population mean confidence interval is $(-6.233^{\circ}\text{C}, -4.718^{\circ}\text{C})$. They used a 90% confidence level.
 - a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
 - b) How much lower could the population mean from population 1 than the population mean from population 2?
 - c) Write the two-population confidence interval sentence explaining this confidence interval.
11. A two-population proportion confidence interval is $(-0.071, +0.068)$. They used a 95% confidence level.
 - a) Is the percentage from population 1 significantly higher, significantly lower, or not significantly different from the percentage from population 2? Explain how you know.
 - b) Write the two-population confidence interval sentence explaining this confidence interval.
12. A two-population mean confidence interval is $(+32.8 \text{ cm}, +37.1 \text{ cm})$. They used a 99% confidence level.
 - a) Is the mean from population 1 significantly higher, significantly lower, or not significantly different from the mean from population 2? Explain how you know.
 - b) How much higher could the population mean from population 1 than the population mean from population 2?
 - c) Write the two-population confidence interval sentence explaining this confidence interval.

(#13-20) Directions: Use the following Statcato and StatKey printouts and answer the following questions.

- a) Does the data meet the assumptions for inference with two population proportions or two population means? If it is two means, are the groups independent or matched pair? List the assumptions needed and how the problem meets them or does not meet them.
- b) Give the sample means or sample proportions for the two groups. Are they close or significantly different? Explain how you know. If they are significantly different, which group has a significantly higher sample mean or sample proportion?



c) Does the confidence interval indicate that the mean or percentage from population 1 is higher, lower, or not significantly different from population 2? Explain how you know. If the mean or percentage from population 1 is higher than population 2, then how much higher could it be? If the mean or percentage from population 1 is lower than population 2, then how much lower could it be?

d) Write the two-population confidence interval sentence explaining this confidence interval.

13. The ACT exam is used by many colleges to test the readiness of high school students for college. Many high school students are now taking ACT prep classes. A local high school offers an ACT prep class, but wants to know if it really helps. Twenty-eight students were randomly selected. They took the ACT exam before and after taking the ACT prep class. Population 1 is the ACT scores after taking the prep class and population 2 is the ACT scores before taking the prep class. The sample mean of the differences was 5.8 ACT points and the sample standard deviation of the differences was 4.3 ACT points. A histogram of the differences was normal. We created a 90% confidence interval for matched pairs with Statcato.

Confidence Interval - Matched Pairs: confidence level = 0.9

Input: Summary data

Difference of Matched Pairs -

N	Mean	Stdev	Margin of Error	90.0%CI
28	5.8	4.3	1.384	(4.4159, 7.1841)

14. We want to compare the population percentage of women that have at least one tattoo (π_1) and the population percentage of men that have at least one tattoo (π_2). A random sample of 794 women found that 137 of them had at least one tattoo. A random sample of 857 men found that 146 of them had at least one tattoo. Go to www.lock5stat.com and use StatKey to create a 99% two-population proportion bootstrap confidence interval.

15. Cotinine is an alkaloid found in tobacco and is used as a biomarker for exposure to cigarette smoke. It is especially useful in examining a person's exposure to second hand smoke. A random sample of 90 non-smoking American adults was collected. These adults were not smokers and did not live with any smokers. The average cotinine level for this sample was 7.2 ng/mL with a standard deviation of 5.8 ng/mL. A second sample of 85 non-smoking American adults was then collected. These adults did not smoke themselves, but did live with one or more smokers. The average cotinine level for this sample was 28.5 and had a standard deviation of 11.4. Population 1 was people that do NOT live with smokers (μ_1) and population 2 was people that DO live with smokers (μ_2). We used Statcato to create the following 95% two-population mean confidence interval for independent groups.

Confidence Intervals - Two population means: confidence level = 0.95

	N	Mean	Stdev
Population 1	90	7.2	5.8
Population 2	85	28.5	11.4

* Population standard deviations are unknown. *

DOF = 123

Margin of error = 2.730

95.0%CI = (-24.0304, -18.5696)



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/18

16. A body mass index of 20-25 indicates that a person is of normal weight. Use the following 90% two-population proportion confidence interval to compare the percentage of men with a normal BMI (π_1) and the percentage of women with a normal BMI (π_2). A random sample of 745 women and 760 men found that 198 of the women and 273 of the men had a normal BMI score.

Confidence Interval - Two population proportions: confidence level = 0.9

	Number of Events	Number of trials	Proportion
Sample 1	273	760	0.359
Sample 2	198	745	0.266

Sample proportion difference = 0.093

Margin of error = 0.039

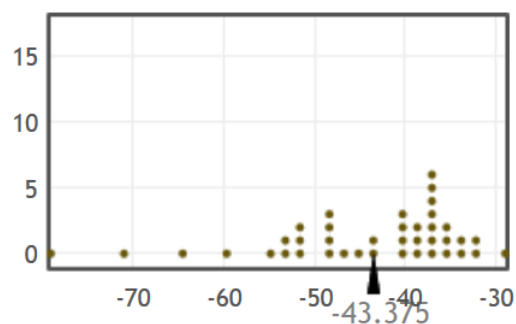
90.0%CI = (0.0543, 0.1325)

17. We used the random health data at www.matt-teachout.org to compare the population mean average systolic and diastolic blood pressures for women. Population 1 was women's diastolic blood pressure and population 2 was women's systolic blood pressure. We used StatKey to create the following 95% bootstrap confidence interval of the differences between the matched pairs.

Original Sample

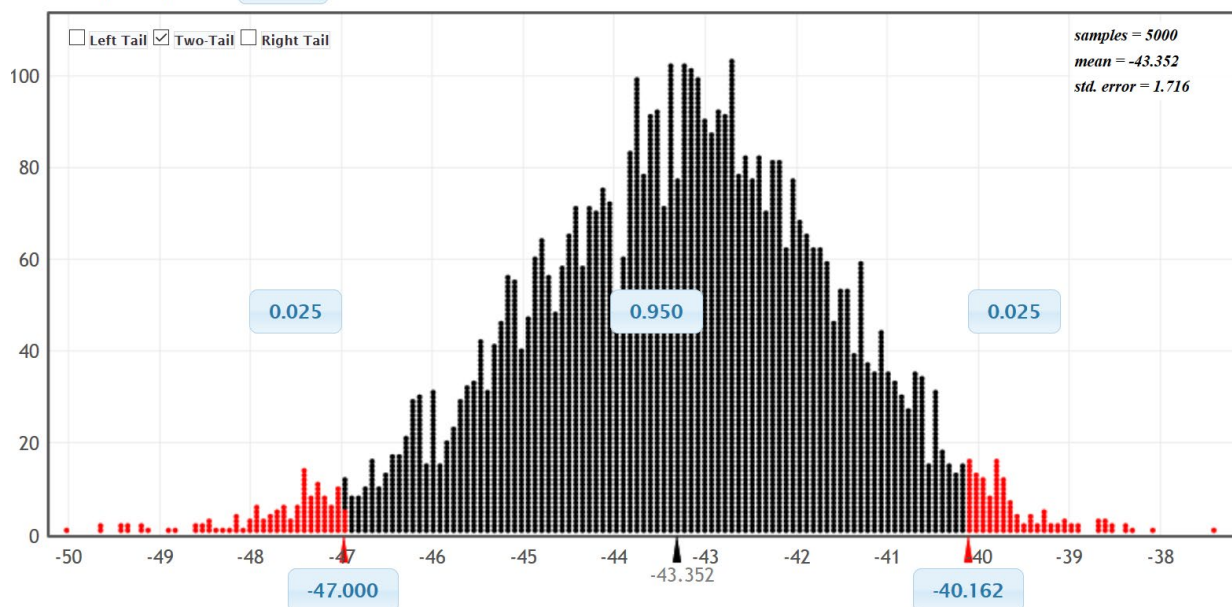
$n = 40$, mean = -43.375

median = -40, stdev = 10.748



Bootstrap Dotplot of

Mean



18. An experiment was done to test the effectiveness of medicine that lowers cholesterol. An experiment was conducted and adults were randomly assigned into two groups. The groups had similar gender, ages, exercise patterns and diet. Of the 410 adults in the treatment group, 49 of them showed a decrease in cholesterol. Of the 420 adults in the placebo group, 38 of them showed a decrease in cholesterol. Was the medicine effective in lowering cholesterol? Use the following 99% confidence interval from Statcato to determine if the percentage of people on the medicine that have a decrease in cholesterol (population 1) is higher than the percentage from the placebo group (population 2).

Confidence Interval - Two population proportions: confidence level = 0.99

	Number of Events	Number of trials	Proportion
Sample 1	49	410	0.120
Sample 2	38	420	0.090

Sample proportion difference = 0.029

Margin of error = 0.055

99.0%CI = (-0.0258, 0.0838)

19. Open the Health data at www.matt-teachout.org. Copy and paste the gender data and cholesterol data into a new excel spreadsheet so that they are next to each other. Go to www.lock5stat.com can click on StatKey. Under the Bootstrap Confidence Interval menu, click on "CI for Difference in Means". Under the "edit data" menu, copy and paste the gender and cholesterol data into StatKey. Construct a 95% two-population mean bootstrap confidence interval estimate of the difference between women's population mean average cholesterol (μ_1) and men's population mean average cholesterol (μ_2).



20. In March 2003, a research group asked 2400 randomly selected Americans whether they believe that the U.S. made the right or wrong decision to use military force in Iraq. Of the 2400 adults, 1862 said that they believed that the U.S. did make the correct decision. In February 2008, the question was asked again to 2180 randomly selected Americans and 684 of them said that the U.S. did make the correct decision. Go to www.lock5stat.com and use StatKey to create a 90% two-population proportion bootstrap confidence interval to compare the population percentage of people that agree with war in 2008 (π_1) and the population percentage in 2003 (π_2).

Section 2G – One-Population Variance & Standard Deviation Confidence Intervals

It is often vital to estimate the standard deviation of a population. However, it can be very difficult to estimate with any accuracy, especially if we only have one random sample. Remember our principle of sampling variability. We saw in previous sections that sample standard deviations (s) will usually be very different from each other and can be very different from the population standard deviation (σ).

One-Population Variance and Standard Deviation Confidence Intervals

Recall that the population standard deviation is the square root of the population variance (σ^2). So we often estimate the population variance and then simply take the square root of the variance to get the standard deviation. The principle of sampling variability also applies to variance. Sample variances (s^2) will usually be very different from each other and may be very different from the population variance (σ^2).

Sampling distributions for variance are usually skewed to the right and rarely have a normal shape. Since the sampling distribution is not normal or symmetric, we cannot use the traditional formula approach of the sample statistic \pm margin of error. That formula will not work.

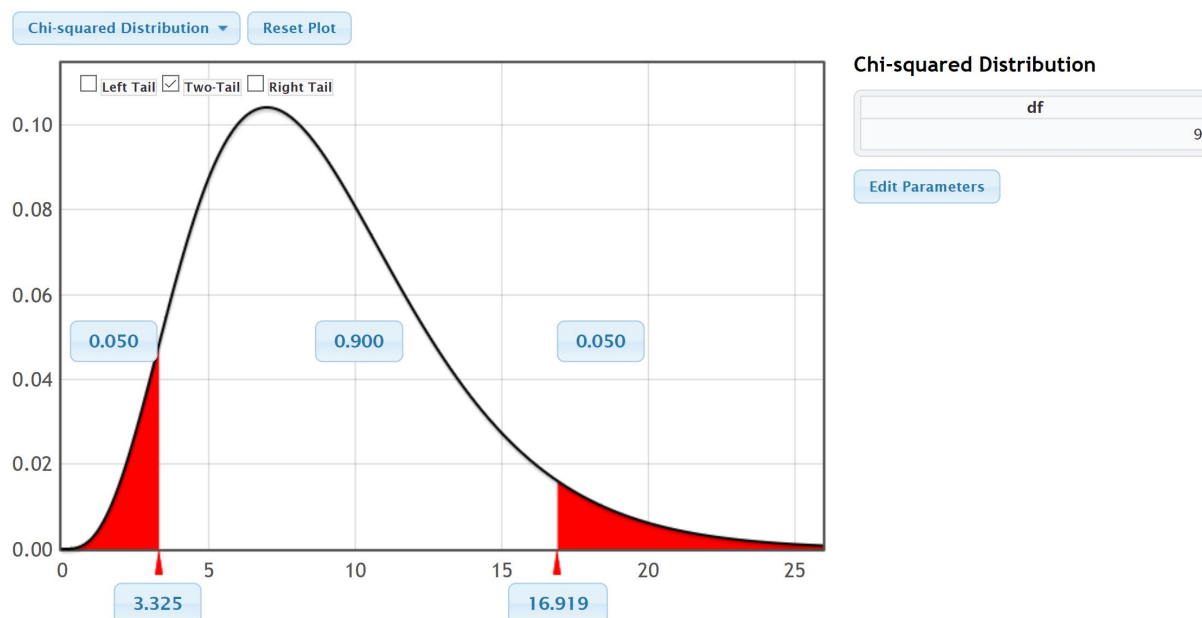
It is important to note that even though the quantitative data itself may be normal, the sampling distribution for variance still may be skewed to the right. Statisticians discovered that as long as the quantitative data itself was normal, the sampling distribution for sample variance follows a Chi-Squared distribution with degrees of freedom ($df = n - 1$). So the formula for making a confidence interval to estimate population variance uses Chi-Squared critical values (χ^2). It is important to note that no matter what the sample size is, the sample data must be normal for this formula to work. If the data is not normal, we must resort to another technique like bootstrapping.

Calculating Chi-Squared Critical Values

Example 1: Calculate the Chi-Squared (χ^2) critical values for a sample size $n = 10$ and a 90% confidence level.

Go to www.lock5stat.com and click on “StatKey”. Under the “theoretical distributions” menu, click on “ χ^2 ”. Since the sample size is 10, the degrees of freedom will be $df = 10 - 1 = 9$. If we click on “two tail” and set the middle proportion to 0.90, we will get the following. Variance is calculated with a sum of squares. That makes it impossible to ever be negative. That also means the upper and lower Chi-Squared critical values will be very different. The upper Chi-Squared critical value will be the larger number on the right and the lower Chi-Squared critical value will be the smaller number on the left. Both will be positive. You can see that Chi-Squared looks skewed to the right for nine degrees of freedom.

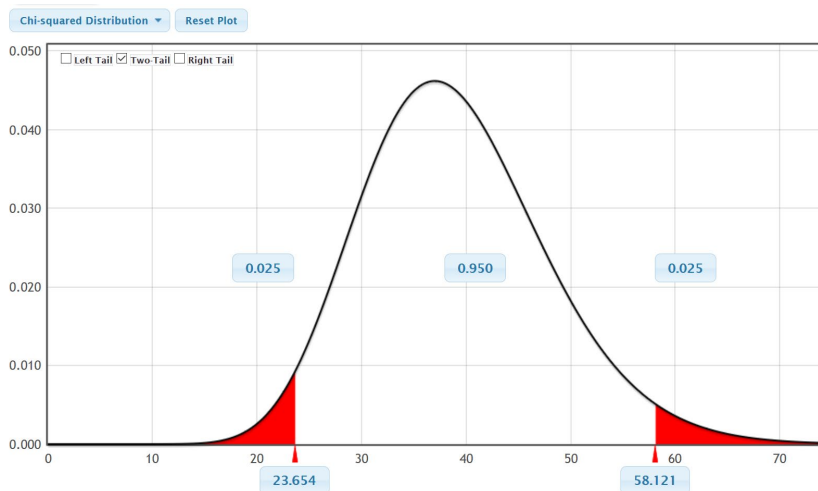




We see from the graph that upper critical value for 90% confidence and 9 degrees of freedom is 16.919 and the lower critical value for 90% confidence and 9 degrees of freedom is 3.325.

Example 2: Calculate the Chi-Squared (χ^2) critical values for a sample size $n = 40$ and a 95% confidence level.

Go to www.lock5stat.com and click on "StatKey". Under the "theoretical distributions" menu, click on " χ^2 ". Since the sample size is 40, the degrees of freedom will be $df = 40 - 1 = 39$. If we click on "two tail" and set the middle proportion to 0.95, we will get the following. Notice that the upper and lower Chi-Squared critical values will both be positive, but will be very different. Also, notice that as the degrees of freedom increases, the chi-squared distribution looks less skewed to the right.



We see from the graph that upper critical value for 95% confidence and 39 degrees of freedom is 58.121 and the lower critical value for 95% confidence and 39 degrees of freedom is 23.654.



Confidence Interval Formulas for Variance and Standard Deviation

Here is the confidence interval formulas for estimating population variance. Taking the square root gives us the formula for estimating population standard deviation as well. Notice that the upper critical value is on the left and lower critical value is on the right. When you divide by a larger number, the overall fraction is smaller.

$$\frac{s^2(n-1)}{\chi^2_{upper}} < \text{Population Variance } (\sigma^2) < \frac{s^2(n-1)}{\chi^2_{lower}}$$

$$\sqrt{\frac{s^2(n-1)}{\chi^2_{upper}}} < \text{Population Standard Deviation } (\sigma) < \sqrt{\frac{s^2(n-1)}{\chi^2_{lower}}}$$

Example 1: We measured the heights in inches of 40 randomly selected men. The data showed a normal shape. The sample standard deviation was 3.020 inches. Use the Chi-squared critical values and the formulas above to create a 95% confidence interval for the population variance and the population standard deviation.

We calculated the Chi-squared critical values in the previous example. The upper critical value was 58.121 and the lower critical value was 23.654.

$$\text{Sample Variance } (s^2) = (3.020)^2 = 9.1204$$

$$\text{Degrees of Freedom } (n - 1) = 40 - 1 = 39$$

$$\frac{s^2(n-1)}{\chi^2_{upper}} < \text{Population Variance } (\sigma^2) < \frac{s^2(n-1)}{\chi^2_{lower}}$$

$$\frac{9.1204 (40-1)}{58.121} < \text{Population Variance } (\sigma^2) < \frac{9.1204 (40-1)}{23.654}$$

$$\frac{9.1204 (39)}{58.121} < \text{Population Variance } (\sigma^2) < \frac{9.1204 (39)}{23.654}$$

$$6.11992 < \text{Population Variance } (\sigma^2) < 15.03744$$

Variance Confidence Interval Sentence: We are 95% confident that the population variance for all men is between 6.11992 and 15.03744 square inches. (*Notice that variance is in square units since it is the standard deviation squared.*)

If we take the square root of our answers, we can get an estimate of the population standard deviation.

$$\sqrt{6.11992} < \text{Population Standard Deviation } (\sigma) < \sqrt{15.03744}$$

$$2.47 \text{ inches} < \text{Population Standard Deviation } (\sigma) < 3.88 \text{ inches}$$

Standard Deviation Confidence Interval Sentence: We are 95% confident that the population standard deviation for all men is between 2.47 inches and 3.88 inches.

As with all calculations, it is much easier and more accurate to calculate these with a computer program.

In Statcato, we can go to the “Statistics” menu and click on confidence intervals. If we click on “1-population variance” and enter the sample size (40) and sample standard deviation 3.020 under summary data, we get the following. Notice you can also calculate the confidence interval from raw data or by entering the sample variance.



1-Population Variance

Help F1

Inputs

☐ Samples in column:

☒ Summarized sample data:

Sample Size: 40

☐ Variance:

☒ Standard deviation: 3.020

Confidence

Confidence level: 0.95 0 - 1.00 (e.g. 0.95)

OK Cancel

Notice Statcato gave us almost the same confidence intervals for variance and standard deviation as we calculated with the formula.

Confidence Interval - One population variance: confidence level = 0.95

Input: Summary data

N	Variance	Stdev	95.0%CI Variance	95.0%CI Stdev
40	9.120	3.02	(6.1200, 15.0372)	(2.4739, 3.8778)

Here are the assumptions for making a confidence interval to estimate population variance or standard deviation.

One-Population Variance or Standard Deviation Confidence Interval Assumptions

1. The quantitative sample data should be collected randomly or be representative of the population.
2. Data values within the sample should be independent of each other.
3. The sample data must be normal.

Does the men's height data meet these assumptions? Let us check them.

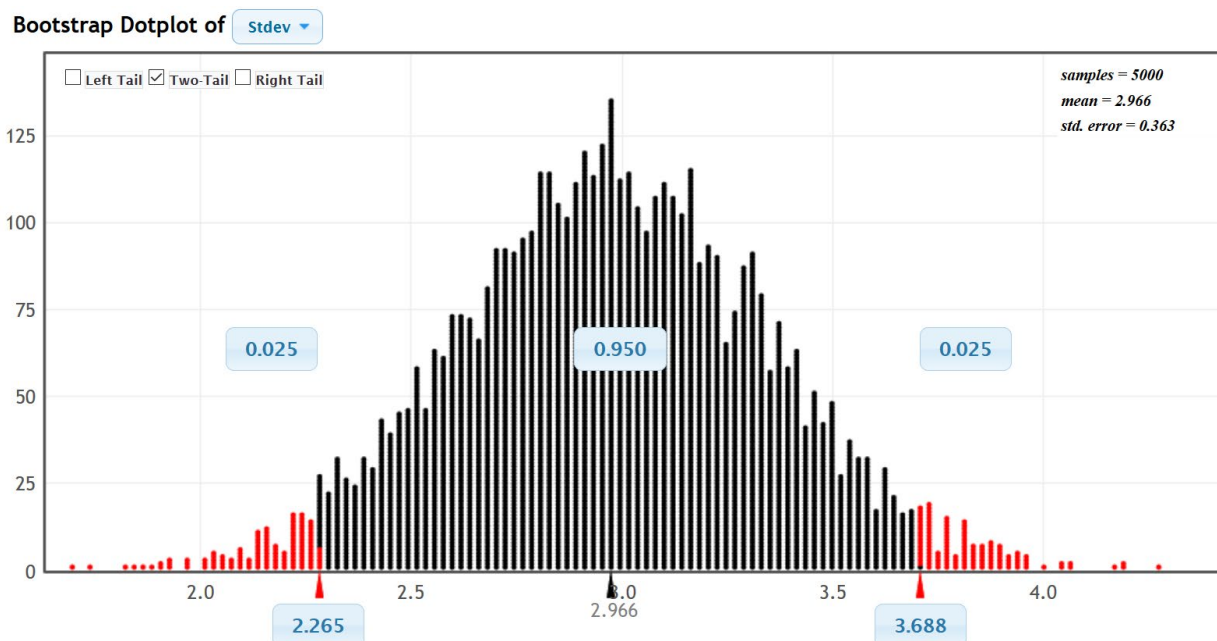
1. Is this random quantitative sample data or sample data that represents the population? Yes. Height is quantitative and this was a random sample.
2. Are the data values within the sample independent of each other? Yes. A random sample out of a large population will be unlikely to accidentally get men that are family members. One man's height should not change the probability of another man's height.
3. Is the sample data normal? We did not see the histogram, but the problem did state that the data was normal.



This chapter is from *Introduction to Statistics for Community College Students*,
 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
 under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Men's Height Bootstrap Example

Let us use a bootstrap distribution to estimate the confidence interval for population standard deviation for men's height. First go to the "Health Data" at www.matt-teachout.org and copy the men's height column of data. Now go to the "Bootstrap Confidence Interval" menu in StatKey at www.lock5stat.com and click on "CI for Single Mean, Median, St.Dev." Under "Edit Data", paste in the raw quantitative men's height data. Make sure to check the "Header Row" box since this data set had a title and push "OK". Under the "Generate Dot plot of" menu, click on "St.Dev." (standard deviation). Now click "Generate 1000 Samples" a few times. Then click "Two-Tail". Make sure the middle proportion is 0.95 (95%).



Notice that the upper and lower limits of the confidence interval are close to what we got with formula or Statcato.

Problems Section 2G

1. What assumptions should we check if we want to use sample data to create a confidence interval to estimate a population standard deviation or variance?
2. A random sample of 45 high school students ACT exams has a skewed left distribution with a sample standard deviation (s) of 9.868. What is the sample variance (s^2)? What is the degrees of freedom? Open StatKey at www.lock5stat.com, go to "Theoretical Distributions" and then click on " χ^2 ". Look up the Chi-squared critical values. Use the critical values, degrees of freedom ($n-1$) and sample variance to construct a 90% confidence interval estimate of the population variance for all ACT exams. Take the square root of your variance confidence interval to calculate a 90% confidence interval estimate for the population standard deviation.

Variance Confidence Interval Formula:

$$\frac{s^2(n-1)}{\chi^2_{upper}} < \text{Population Variance } (\sigma^2) < \frac{s^2(n-1)}{\chi^2_{lower}}$$

a) Sample Variance = (Sample Standard Deviation)² =

b) Degrees of Freedom = $n - 1$ =



This chapter is from *Introduction to Statistics for Community College Students*,
 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
 under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

- c) Chi-squared upper critical value =
- d) Chi-squared lower critical value =
- e) Variance Confidence Interval =
- f) Standard Deviation Confidence Interval =

3. Use the following Statcato printout to check your variance confidence interval answer and your standard deviation confidence interval answer from the random sample ACT data in number 2. Check the assumptions for a variance confidence interval. Remember the data was skewed left. Write down a sentence to explain the population variance confidence interval. Write down a sentence to explain the population standard deviation confidence intervals.

Confidence Interval - One population variance: confidence level = 0.9

Input: Summary data

N	Variance	Stdev	90.0%CI Variance	90.0%CI Stdev
45	97.377	9.868	(70.8423, 143.8391)	(8.4168, 11.9933)

- a) Check each of the assumptions for this problem. Explain your answers.
- b) Write down a sentence to explain the population variance confidence interval.
- c) Write down a sentence to explain the population standard deviation confidence interval.

4. A random sample of body temperatures in degrees Fahrenheit was taken from 50 randomly selected adults. Assume the sample data was normally distributed. The sample standard deviation of 0.765 °F. What is the sample variance (s^2)? What is the degrees of freedom? Open StatKey at www.lock5stat.com, go to "Theoretical Distributions" and then click on " χ^2 ". Look up the Chi-squared critical values. Use the critical values, degrees of freedom ($n-1$) and sample variance to construct a 99% confidence interval estimate of the population variance for all body temperatures. Take the square root of your variance confidence interval to calculate a 99% confidence interval estimate for the population standard deviation.

Variance Confidence Interval Formula:

$$\frac{s^2(n-1)}{\chi^2_{upper}} < \text{Population Variance } (\sigma^2) < \frac{s^2(n-1)}{\chi^2_{lower}}$$

- a) Sample Variance = (Sample Standard Deviation)² =
- b) Degrees of Freedom = $n - 1$ =
- c) Chi-squared upper critical value =
- d) Chi-squared lower critical value =
- e) Variance Confidence Interval =
- f) Standard Deviation Confidence Interval =



5. Use the following Statcato printout to check your variance confidence interval answer and your standard deviation confidence interval answer from the random sample body temperature data in number 4. Check the assumptions for a variance confidence interval. Remember the data is normally distributed. Write down a sentence to explain the population variance confidence interval. Write down a sentence to explain the population standard deviation confidence intervals.

Confidence Interval - One population variance: confidence level = 0.99

Input: Summary data

N	Variance	Stdev	99.0%CI Variance	99.0%CI Stdev
50	0.585	0.765	(0.3666, 1.0524)	(0.6054, 1.0258)

- Check each of the assumptions for this problem. Explain your answers.
- Write down a sentence to explain the population variance confidence interval.
- Write down a sentence to explain the population standard deviation confidence interval.

6. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "sugar (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Stdev". Now click on "Edit Data" and paste the sugar data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 99% confidence interval for the population standard deviation.

- Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.
- How many bootstrap samples did you take?
- What is the shape of the bootstrap distribution for the standard deviation?
- Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.
- Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.

7. Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the "cereal data" in excel. Copy the column of data labeled "carbs (grams per serving)". Go to www.lock5stat.com and click on the "StatKey" tab. Under the "Bootstrap Confidence Intervals" menu, click on "CI for Single Mean, Median, St.Dev." Click on "Bootstrap Dot plot of Stdev". Now click on "Edit Data" and paste the carb data into StatKey. Click on "Generate 1000 Samples" a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 95% confidence interval for the population standard deviation.

- Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.
- How many bootstrap samples did you take?
- What is the shape of the bootstrap distribution for the standard deviation?
- Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.
- Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.



8. Go to www.matt-teachout.org, click on “Statistics” and then “Data Sets”. Open the “bear data” in excel. Copy the column of data labeled “weight in pounds”. Go to www.lock5stat.com and click on the “StatKey” tab. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Single Mean, Median, St.Dev.” Click on “Bootstrap Dot plot of Stdev”. Now click on “Edit Data” and paste the bear weight data into StatKey. Click on “Generate 1000 Samples” a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 90% confidence interval for the population standard deviation for the weight of bears.

- Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.
- How many bootstrap samples did you take?
- What is the shape of the bootstrap distribution for the standard deviation?
- Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.
- Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.

9. Go to www.matt-teachout.org, click on “Statistics” and then “Data Sets”. Open the “bear data” in excel. Copy the column of data labeled “length in inches”. Do not click on “head length” by mistake. We want the overall length of the bears. Go to www.lock5stat.com and click on the “StatKey” tab. Under the “Bootstrap Confidence Intervals” menu, click on “CI for Single Mean, Median, St.Dev.” Click on “Bootstrap Dot plot of Stdev”. Now click on “Edit Data” and paste the bear length data into StatKey. Click on “Generate 1000 Samples” a few times to create the bootstrap sampling distribution for the standard deviation. Use the bootstrap distribution to find a 99% confidence interval for the population standard deviation for the length of bears.

- Does this data meet the assumptions for a bootstrap confidence interval? Assume the data was collected randomly. Explain your answer.
 - How many bootstrap samples did you take?
 - What is the shape of the bootstrap distribution for the standard deviation?
 - Write the upper and lower limits of the bootstrap confidence interval for the population standard deviation.
 - Write a sentence to explain the bootstrap confidence interval estimate of the population standard deviation.
-

Chapter 2 Review Problems

Topics to Study

- Confidence Interval Key Terms
- Statistics and Parameters
- Sampling Distributions
- Know how to interpreting confidence intervals
- T-distribution
- Table summarizing critical value, standard error, margin of error and confidence intervals
- Confidence Interval Assumptions
- Bootstrapping
- Two-population confidence intervals



This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

1. Determine if each of the following symbols are a mean, standard deviation, proportion, slope, or correlation coefficient. Also, decide if it is a sample statistic or a population parameter.

(N , n , π , \hat{p} , μ , \bar{x} , σ , s , ρ , r , β_1 , b_1 , σ^2 , s^2)

2. For each number determine the symbol used from the following list and if it is a statistic or a parameter.

(N , n , π , \hat{p} , μ , \bar{x} , σ , s , ρ , r , β_1 , b_1 , σ^2 , s^2)

- "We tested a sample of incoming college freshman and found that their sample mean average IQ was 101.9, a sample standard deviation of 14.8 and a sample variance of 219.04. We think the population mean IQ is 100, the population standard deviation for IQ scores is 15, and the population variance is 225."
- "We want to check and see if the population correlation coefficient could be zero and the population slope could be about 20 pounds per degree Fahrenheit. The sample correlation coefficient was 0.338 and the sample slope was 13.79 pounds per degree Fahrenheit."
- "Our study found that of the people tested in the sample, only 3% showed side effects to the medication. We think the population percentage of side effects is closer to 1.5%".
- "We took a random sample of 238 people from a population of about 5 million people."

3. List the assumptions that need to be checked before you make a one-population mean confidence interval.

4. List the assumptions that need to be checked before you make a one-population variance or standard deviation confidence interval.

5. List the assumptions that need to be checked before you make a one-population proportion confidence interval.

6. List the assumptions that need to be checked before you make a two-population mean confidence interval.

7. List the assumptions that need to be checked before you make a two-population proportion confidence interval.

8. List the assumptions for a bootstrap confidence interval.

9. Define the following terms: Population, Census, Sample, Statistic, Parameter, Sampling Distribution, Sampling Variability, Point Estimate, Margin of Error, Standard Error, Confidence Interval, 95% Confident, 90% Confident, 99% Confident, Bootstrapping, Bootstrap Sample, Bootstrap Statistic, Bootstrap Distribution

10. Write a sentence to explain the following confidence intervals. Assume the confidence intervals came from unbiased random sample data that met all of the assumptions.

- Explain the one-population mean confidence interval (55.6 pounds, 69.4 pounds).
Confidence Level = 99%
- Explain the one-population proportion confidence interval (0.352, 0.411). *Confidence Level = 90%*
- Explain the one-population standard deviation confidence interval (3.1 pounds, 4.7 pounds).
Confidence Level = 95%
- Explain the one-population variance confidence interval (461.8 square inches, 591.3 square inches).
Confidence Level = 99%
- Explain the two-population mean confidence interval (+13.2 kg, +14.8 kg). *Confidence Level = 95%*
Is there a significant difference between the populations? Explain why.
- Explain the two-population mean confidence interval (-\$3.79, +\$4.13). *Confidence Level = 90%*
Is there a significant difference between the populations? Explain why.



- g) Explain the two-population proportion confidence interval $(-0.024, +0.017)$. *Confidence Level = 95%*
Is there a significant difference between the populations? Explain why.
- h) Explain the two-population proportion confidence interval $(-0.072, -0.057)$. *Confidence Level = 99%*
Is there a significant difference between the populations? Explain why.
11. Explain what a sampling distribution is and how we can use it to find the population parameter, standard error and better understand sampling variability.
12. Explain how a critical value Z-score or T-score and standard error can be used to calculate the margin of error. How can we use margin of error to make the confidence interval.
13. In one-population variance confidence intervals, how does the computer use the chi-squared critical values, the degrees of freedom and the sample variance to make the confidence interval?
14. Answer the following questions about the T-distribution.
- Who invented the T-distribution?
 - What company did he work for?
 - Why did he invent T-scores?
 - Why did he have to publish the T-distribution discovery under a pseudonym?
 - What pseudonym did he use?
 - When are T-scores significantly larger than Z-scores?
 - When are T-scores and Z-scores almost the same?
 - What types of confidence intervals use Z-scores?
 - What types of confidence intervals use T-scores?
 - How is degrees of freedom usually calculated for one quantitative data set?
15. Explain the ideas behind the Central Limit Theorem.
16. Explain the process of bootstrapping and how a bootstrap distribution may be used to calculate a confidence interval without a formula. What assumptions are necessary to make a bootstrap confidence interval? How is a bootstrap distribution different from a sampling distribution?
-

Chapter 3: Introduction to Hypothesis Testing

Vocabulary

Population: The collection of all people or objects to be studied.

Sample: Collecting data from a small subgroup of the population.

Random Sample: Sample data collected in such a way that everyone in the population has an equal chance to be included.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about "no change", "no relationship" or "no effect".



*This chapter is from Introduction to Statistics for Community College Students,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Alternative Hypothesis (H_A or H_1): A statement about the population that does not involve equality. It is often a statement about a “significant difference”, “significant change”, “relationship” or “effect”.

Population Claim: What someone thinks is true about a population.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

One-Population Proportion Test Statistic (z): The sample proportion is this many standard errors above or below the population proportion in the null hypothesis.

One-Population Mean Test Statistic (t): The sample mean is this many standard errors above or below the population mean in the null hypothesis.

Critical Value: A number we compare our test statistic to in order to determine significance. In a sampling distribution or a theoretical distribution approximating the sampling distribution, the critical value shows us where the tail or tails are. The test statistic must fall in the tail to be significant.

Sampling Variability: Also called “random chance”. The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and sampling variability. If the P-value is lower than the significance level, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened because of sampling variability.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value and standard error without a formula.

Type 1 Error: When biased sample data leads you to support the alternative hypothesis when the alternative hypothesis is actually wrong in the population.

Type 2 Error: When biased sample data leads you fail to reject the null hypothesis when the null hypothesis is actually wrong in the population.

Beta Level (β): The probability of making a type 2 error.

Conclusion: A final statement in a hypothesis test that addresses the claim and evidence.

Introduction: The goal of statistics is to learn about the world around us. While we may collect sample data, our goal is not just to analyze sample data. We want to know what is happening in the population. Sometimes, people make guesses about what they think is happening in the population. These guesses are often called “claims”. In this chapter, we will discuss the process of hypothesis testing. A hypothesis test is a scientific procedure for using representative random sample data to investigate claims about populations. A hypothesis test involves many difficult concepts and has many steps. It is difficult to learn all of hypothesis testing at once. For this reason, we will be learning the steps for hypothesis testing one at a time. Eventually we will put all of the parts together and complete the hypothesis test from start to finish.



Section 3A – Null and Alternative Hypothesis

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about “no change”, “no relationship” or “no effect”.

Alternative Hypothesis (H_A or H_1): A statement about the population that does not involve equality. It is often a statement about a “significant difference”, “significant change”, “relationship” or “effect”.

Population Claim: What someone thinks is true about a population.

Introduction:

Remember, a hypothesis test is a procedure for checking what someone has said about the population. This is often called the population “claim”. From this claim, we will need to identify two opposing views. The claim and the rival hypothesis. If the claim is not true, then what would be true? These two opposing views are referred to as the null hypothesis (H_0) and the alternative hypothesis (H_A or H_1). The symbol often used for “null hypothesis” is “ H_0 ”. The symbol used to represent the alternative hypothesis is “ H_A ” or “ H_1 ”. I prefer “ H_A ” for alternative hypothesis.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about “no change”, “no relationship” or “no effect”.

Alternative Hypothesis (H_A): A statement about the population that does not involve equality. It is often a statement about a “significant difference”, “significant change”, “relationship” or “effect”.

Important Notes:

- In the last chapter, we talked about symbols that represent population parameters like the population mean (μ), the population proportion (π) or the population standard deviation (σ). The null and alternative hypothesis are competing ideas about the population and only involve population parameters like μ , π or σ . The sample data and sample statistics like the sample proportion (\hat{p}), the sample mean (\bar{x}) or the sample standard deviation (s) are never part of the null or alternative hypotheses.

A hypothesis test is a procedure for deciding between two opposing views about the population. Sometimes the person will tell you the two opposing views, but there will be one view that the person thinks is true or wants you as the data scientist to give evidence toward. This is called the “claim”.

Claim: What the person now thinks is true about the population. The claim can be a question that someone needs to figure out. It can also be an opinion about the population that they want you to investigate.



*This chapter is from **Introduction to Statistics for Community College Students**,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Let's go over some basic steps to writing the null and alternative hypothesis.

Steps for finding the Null and Alternative Hypothesis

1. Write down the two competing views about the population in symbolic language. Make sure to determine if it is one-population, or two-population, or more and the correct letter (parameter) to use. If the person only gives you his or her claim, we will often use the opposite as the rival hypothesis or competing view.
2. Write the word "claim" next to what the person thinks is true or what they are asking you to provide evidence toward.
3. The statement that has " $=$ " or " \geq " or " \leq " is the null hypothesis. It is usually " $=$ ". Put an " H_0 " next to it. Remember, the null hypothesis is usually a statement about no change, no effect, or not related. That is why the null hypothesis is often given with " $=$ ".
4. The statement that has " \neq " or " $<$ " or " $>$ " is the alternative hypothesis. Put an " H_A " next to it. This is usually a statement about something changing or being related or having a significant effect.
5. Determine what type of test you are dealing with. Is it right-tailed, left-tailed or two-tailed? In a hypothesis test, we are often attempting to "reject the null hypothesis". This would happen if we believed the alternative hypothesis is correct. For this reason, the alternative hypothesis determines what type of test you are doing. In a one-population or two-population test, if H_A is less than ($<$), it is a left tailed test. (Notice " $<$ " points to the left). If H_A is greater than ($>$), it is a right tailed test. (Notice " $>$ " points to the right). If H_A is not equal (\neq), it is a two-tailed test. (not equal means less than or greater than.) Things get more complicated if we are dealing with 3 or more populations. We will deal with these cases in the next chapter.
 - Left-tailed test: H_A is less than ($<$)
 - Right-tailed test: H_A is greater than ($>$)
 - Two-tailed test: H_A is less than (\neq)

Symbols for population parameters:

μ (population mean)

π or p (population proportion/percentage)

σ (population standard deviation)

σ^2 (population variance)

Important Notes

- Never use a sample statistic (\bar{x}, \hat{p}, s). Remember a null and alternative hypothesis are statements about the population parameters (μ, π, σ).
- The symbol " $<$ " means less than. Notice the symbol for "less than" points to the left. The opposite of less than is " \geq " (greater than or equal to).
- The symbol " $>$ " means less than. Notice the symbol for "greater than" points to the right. The opposite of less than is " \leq " (less than or equal to).
- For one-population, always write the letter on the left side and the number on the right side. $\mu < 50$ (never as $50 > \mu$)
- For two-population, always put population 1 on the left side and population 2 on the right side. $\pi_1 > \pi_2$ (never as $\pi_2 < \pi_1$).



Three Types of Hypothesis Tests

- Hypothesis Tests are designated as one of three types. It is important to know what type of test you are doing.
- Note that the alternative hypothesis H_A decides the type of test. In a hypothesis test, we are often attempting to “reject the null hypothesis”. This would happen if we believed the alternative hypothesis is correct. For this reason, the alternative hypothesis determines what type of test you are doing. In a one-population or two-population test, if H_A is less than ($<$), it is a left tailed test. (Notice “ $<$ ” points to the left). If H_A is greater than ($>$), it is a right tailed test. (Notice “ $>$ ” points to the right). If H_A is not equal (\neq), it is a two-tailed test. (not equal means less than or greater than.)

Left-tailed test: H_A is less than ($<$) (points to the left)

Right-tailed test: H_A is greater than ($>$) (points to the right)

Two-tailed test: H_A is less than (\neq)

Note: If you have more than two populations, things become more complicated. For example, ten populations will often be condensed into one right-tailed test. We will explain more about these tests in the next chapter.

Example 1: Auto Magazine Article

“The population mean average weight of car transmissions used to be about 300 pounds. With more and more small car options, we think the population mean average weight of car transmissions has decreased.”

Step 1: The population views are about one population mean (μ). Notice there are two opposing views are the population mean given. Write down the two opposing views in symbolic language. Notice again that “decreased” means less than “ $<$ ” which points to the left. Make sure to put the symbol on the left and the number on the right. (“ $\mu < 300$ ” NOT “ $300 > \mu$ ”). Notice also that the opposing view is that it used to be exactly 300 pounds. They never believed that the population mean average was more than 300 pounds. It seems more appropriate to use “ $=$ ” as the opposing view instead of “ \geq ”.

$$\mu = 300$$

$$\mu < 300$$

Step 2: Let’s identify the claim. It seems they think that it used to be 300 pounds, but now they think the population mean is really less than 300 pounds. This would be the claim.

$$\mu = 300$$

$$\mu < 300 \text{ (Claim)}$$

Step 3. The statement that has “ $=$ ” or “ \geq ” or “ \leq ” is the null hypothesis. Put an “ H_0 ” next to it. Notice that “claim” has nothing to do with a statement being a null or alternative hypothesis. It is the symbol that decides H_0 .

$$H_0: \mu = 300$$

$$\mu < 300 \text{ (Claim)}$$

Step 4: The statement that has “ \neq ” or “ $<$ ” or “ $>$ ” is the alternative hypothesis. Put an “ H_A ” next to it. Notice that “claim” has nothing to do with a statement being a null or alternative hypothesis. It is the symbol that decides H_A .

$$H_0: \mu = 300$$

$$H_A: \mu < 300 \text{ (Claim)}$$



Step 5: Determine the type of test you are doing. Remember, the alternative hypothesis determines the type of test. If H_A is less than ($<$), it is a left tailed test. (Notice " $<$ " points to the left). If H_A is greater than ($>$), it is a right tailed test. (Notice " $>$ " points to the right). If H_A is not equal (\neq), it is a two-tailed test. (not equal means less than or greater than.)

In this problem, the alternative hypothesis is less than ($<$) which points to the left. So this is a left tailed test.

$$H_0: \mu = 300$$

$$H_A: \mu < 300 \text{ (Claim)}$$

Left Tailed Test

Example 2: Medication Side Effects

"The FDA says that about 2.5% of people that take this medicine will have serious side effects".

Step 1: The population views are about one population proportion (π or p). Notice that only one view about the population proportion is given. In this case we will have to think about opposites. The FDA seems to be making a claim about equaling 2.5% (0.025). They did not specify higher or lower. The opposite of equal is not equal (\neq) so we will use that as our opposing view. Write down the two opposing views in symbolic language. Population percentage claims are usually written as a decimal proportion. Make sure to put the symbol on the left and the number on the right. (" $\pi = 0.025$ " NOT " $0.025 = \pi$ ").

$$\pi = 0.025$$

$$\pi \neq 0.025$$

OR

$$p = 0.025$$

$$p \neq 0.025$$

Step 2: Let's identify the claim. It seems that the FDA thinks it is 2.5%. Don't let the word "about" change your mind about this. Remember, we never know anything definite about millions of people. Language like "about" or "around" is often used in population claims. They did not specify higher or lower, so it is still an equal to claim.

$$\pi = 0.025 \text{ (Claim)}$$

$$\pi \neq 0.025$$

OR

$$p = 0.025 \text{ (Claim)}$$

$$p \neq 0.025$$

Step 3. The statement that has " $=$ " or " \geq " or " \leq " is the null hypothesis. Put an " H_0 " next to it. Notice that "claim" has nothing to do with a statement being a null or alternative hypothesis. It is the symbol that decides H_0 .

$$H_0: \pi = 0.025 \text{ (Claim)}$$

$$\pi \neq 0.025$$

OR

$$H_0: p = 0.025 \text{ (Claim)}$$

$$p \neq 0.025$$



Step 4: The statement that has “ \neq ” or “ $<$ ” or “ $>$ ” is the alternative hypothesis. Put an “ H_A ” next to it. Notice that “claim” has nothing to do with a statement being a null or alternative hypothesis. It is the symbol that decides H_A .

$H_0: \pi = 0.025$ (Claim)

$H_A: \pi \neq 0.025$

OR

$H_0: p = 0.025$ (Claim)

$H_A: p \neq 0.025$

Step 5: Determine the type of test you are doing. Remember, the alternative hypothesis determines the type of test. If H_A is less than ($<$), it is a left tailed test. (Notice “ $<$ ” points to the left). If H_A is greater than ($>$), it is a right tailed test. (Notice “ $>$ ” points to the right). If H_A is not equal (\neq), it is a two-tailed test. (not equal means less than or greater than.)

In this problem, the alternative hypothesis is not equal (\neq), so this is a two-tailed test.

$H_0: \pi = 0.025$ (Claim)

$H_A: \pi \neq 0.025$

OR

$H_0: p = 0.025$ (Claim)

$H_A: p \neq 0.025$

Two-Tailed Test

Example 3: Comparing female and male SAT scores

The school board claims that the average SAT score for female high school students is greater than the average SAT score for male high school students. If gender is not related to SAT scores, then the SAT scores should be the same.

Step 1: Since this is a two-population mean average problem we will need to decide what is population 1 and what is population 2 and the correct letter to use. I tend to write statements as they are claimed. In this problem, they said that the population mean average SAT for females is higher than males. The most straight forward way is to make females population 1 and males population 2 and then say females is higher than males. We could reverse it and put males as population 1, but then we would have to say that population 1 is lower than population 2. Remember that the population 1 parameter should always go on the left.

μ_1 : Female

μ_2 : Male

Write down the two opposing views in symbolic language. Notice that the two opposing views indicate a gender / SAT relationship (the mean average for females is higher than males) verses no relationship (mean averages are the same.) Notice “equal to” goes with not related. This is why not related is always the null hypothesis. Even though the opposite of $>$ is “ \leq ”, this symbol does not seem appropriate for this test. If the female SAT’s were lower, that also would indicate a gender/SAT relationship.

$\mu_1 > \mu_2$

$\mu_1 = \mu_2$

Step 2: Decide the claim. It seems that while the school board longs for the day when the genders are the same, they do not believe that is true right now. They believe that gender does matter. They believe that the population mean average for females is higher than for males.



$$\mu_1 > \mu_2 \text{ (Claim)}$$

$$\mu_1 = \mu_2$$

Step 3: Decide which statement is the null hypothesis. Remember, the statement that has “=” or “≥” or “≤” is the null hypothesis. Put an “ H_0 ” next to it. Remember, claim has nothing to do with it.

$$\mu_1 > \mu_2 \text{ (Claim)}$$

$$H_0: \mu_1 = \mu_2$$

Step 4: Decide which statement is the alternative hypothesis. Remember, the statement that has “≠” or “<” or “>” is always the alternative hypothesis. Put an “ H_A ” next to it. Claim has nothing to do with a statement being the alternative hypothesis. It is the symbol.

$$H_A: \mu_1 > \mu_2 \text{ (Claim)}$$

$$H_0: \mu_1 = \mu_2$$

Step 5: Determine the type of test you are doing. Remember, the alternative hypothesis determines the type of test. If H_A is less than (<), it is a left tailed test. (Notice “<” points to the left). If H_A is greater than (>), it is a right tailed test. (Notice “>” points to the right). If H_A is not equal (≠), it is a two-tailed test. (not equal means less than or greater than.)

In this problem, the alternative hypothesis is greater than (>), which points to the right. So this is a right-tailed test.

Null Hypothesis Confusion

Some students get confused over the sign in the null hypothesis. It is important to pay attention to the language. Look at the following two examples.

Claim: “The population standard deviation (σ) used to be 2 inches but now we think it has increased.”

Since we have two opposing views, we can write them both and then chose the null and alternative hypothesis by the sign.

$$H_0: \sigma = 2$$

$$H_A: \sigma > 2 \text{ (Claim)}$$

Claim: “We think that the population standard deviation is more than 2 inches.”

This time, we do not have the opposing view, so we use opposites. The opposite of “>” is “≤”.

$$H_0: \sigma \leq 2$$

$$H_A: \sigma > 2 \text{ (Claim)}$$

These two examples illustrate a point of confusion. Sometimes you may see the null hypothesis of the same hypothesis test written as “=” and sometimes it may be written with “≤” or “≥”.

$$H_0: \sigma = 2$$

$$H_A: \sigma > 2 \text{ (Claim)}$$

OR

$$H_0: \sigma \leq 2$$

$$H_A: \sigma > 2 \text{ (Claim)}$$



Either answer is ok. Notice the parameter is still 2 inches and they are both right-tailed tests. In all practicality, they are the same test.

Many scientists prefer to write the null almost always as “=”. Remember the null in an experiment is usually “no change” or “no effect”. Change of any kind is usually denoted by the alternative hypothesis.

Confusion about “At Least” or “At Most”.

When we say we have at least \$20, we mean the amount of money is greater than or equal to \$20 (\geq).

When we say we have at most \$20, we mean the amount of money is less than or equal to \$20 (\leq).

If we stick to the language of “at least” or “at most” we would have to make them the null hypothesis.

If the claim was that the population mean amount of money is at least \$20, we would write the following null and alternative hypothesis. Remember at least means greater than or equal to “ \geq ”. The opposing view would be less than “ $<$ ”. Since the alternative is less than, this would be a left tailed test.

$$H_0: \mu \geq 20 \text{ (Claim)}$$

$$H_A: \mu < 20$$

This can create confusion. Does the person really want to check “at least” or do they really mean “more than”? It has been my experience that when people want to check an “at least” claim, they are better off changing the claim to “more than” or “increased” and doing a traditional right tailed test.

Similarly if someone wants to test an “at most” claim, they are better off changing the claim to “less than” or “decreased” and doing a traditional left tailed test.

Practice Problems Section 3A

For each of the following problems:

- a) Write the null and alternative hypothesis.
- b) Label whether the null or the alternative is the claim.
- c) Tell whether this is a left tail test, a right tail test, or a two tail test.

1. According to a CNN report, 93% of all Americans also own a traditional phone. We disagree with this report. We claim that the percentage has decreased as more and more Americans opt to only use a cell phone and throw away their traditional phones.
2. According to a recent Newspaper article, the population mean average amount of time people in California spend eating and drinking per day is 1.25 hours. Test the claim that the population mean average number of hours spent eating and drinking really is 1.25 hours.
3. According to an article in *USA Today*, 74% of Americans own a credit card. We disagree with the *USA Today* article and claim that more than 74% of Americans own a credit card.
4. It has long been thought that the population mean average body temperature for all humans is 98.6 degrees Fahrenheit. A recent scientific study is now claiming that the population mean average body temperature is actually lower than 98.6 degrees Fahrenheit.
5. The population standard deviation for the heights of adult men was thought to be 2.9 inches. We disagree with this. We claim that the population standard deviation for heights of men is not 2.9 inches.
6. Test the claim that more than 10% of the world population is left-handed.



7. The population percent of all women worldwide that hold CEO level jobs is lower than the population percent of all men worldwide that hold CEO level jobs. We long to eliminate this gender bias, so that the population percentages are the same.
 8. The population variance for cholesterol levels is about the same for adult men and adult women in the U.S.
 9. The majority of all Republicans support decreasing taxes.
 10. We claim that the population correlation coefficient is zero.
 11. While some think that the population slope is zero, we claim that the population slope is significantly positive.
 12. According to the center for disease control (CDC), 9.4% of people in the U.S. have some form of diabetes.
 13. According to an article in an automobile magazine, the population mean average price of a used mustang is 18 thousand dollars. We disagree with this article and claim that data indicates that the population mean is lower than 18 thousand dollars.
 14. According to the center for disease control (CDC), the population percentage of rabies cases from wild animals in 2015 is different from the percentage of rabies cases from wild animals in 2017.
 15. Some people believe that the population variance for human body temperature is about 0.5 degrees Fahrenheit squared. A recent scientific study is now claiming that the population variance for body temperature may actually be higher than 0.5.
 16. Test the claim that less than 90% of the world population is right-handed.
 17. We claim that the population mean average salary of female lawyers in New York is significantly lower than the population mean average salary of male lawyers in New York. We long to eliminate this gender bias, so that the population mean averages are the same.
 18. The population variance for men's heights is not the same as for women's heights.
 19. We claim that the population correlation coefficient is greater than zero.
 20. While some think that the population slope is zero, we claim that the population slope is significantly negative.
-

Section 3B – Test Statistics and Critical Values

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about "no change", "no relationship" or "no effect".



*This chapter is from Introduction to Statistics for Community College Students,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

One-Population Proportion Test Statistic (z): The sample proportion is this many standard errors above or below the population proportion in the null hypothesis.

One-Population Mean Test Statistic (t): The sample mean is this many standard errors above or below the population mean in the null hypothesis.

Two-Population Proportion Test Statistic (z): The sample proportion for group one is this many standard errors above or below the sample proportion for group two. Can also be thought of as how many standard errors the difference between the sample proportions is from zero or from some other specific difference.

Two-Population Mean Average Test Statistic (t): The sample mean for group one is this many standard errors above or below the sample mean for group two. Can also be thought of as how many standard errors the difference between the sample means is from zero or from some other specific difference.

Critical Value: A number we compare our test statistic to in order to determine significance. In a sampling distribution or a theoretical distribution approximating the sampling distribution, the critical value shows us where the tail or tails are. The test statistic must fall in the tail to be significant.

Introduction

A hypothesis test compares random sample data to the null and alternative hypothesis in order to determine the validity of a population claim. Remember, the null hypothesis is the population statement that involves equality, usually to some population parameter. The principle of sampling variability (random chance) tells us that random sample statistics will usually be very different than population parameters in the null hypothesis. We know the sample data will almost always disagree with the null hypothesis. The question is does it significantly disagree?

Significant differences are very difficult to determine. For example, a sample mean is 4.3 kg higher than the population mean in the null hypothesis, but is that significant? A sample percentage may be 3.5% below the population percentage in the null hypothesis, but is that significant? Statisticians and mathematicians put a lot of thought into figuring out ways to determine if a sample statistic significantly disagrees with a population parameter. The result was the invention of the "test statistic".

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis.

Data can take on many forms, categorical or quantitative, one-population or two-population or ten-populations. So there are many types of test statistics to deal with the various types of data and the number of populations. Test statistics have one thing in common. They are trying to see if the random sample data significantly disagrees with the null hypothesis or if it is not a significant disagreement.

Using Test Statistics and Critical Values to Judge Significance

Key question: How can we tell if the test statistic indicates a significant disagreement between the sample data and the null hypothesis?

The answer to this is to compare the test statistic to a critical value.

Critical Value: We compare a test statistic to this number to determine if the sample data significantly disagrees with the null hypothesis.



*This chapter is from **Introduction to Statistics for Community College Students**, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18*

Critical values and test statistics are always different depending on the situation. Am I dealing with a left-tailed, right-tailed or two-tailed test? What confidence level (or significance level) should I use? Most computer programs calculate the test statistics and critical values for a situation. Your job as a data scientist is to be able to explain these numbers and judge significance.

The most important rule to remember about test statistics is that the following. You may not understand the test statistic calculation or how the critical value was found, but you can at least judge whether your random sample data significantly disagrees with the null hypothesis or not.

Critical values help us judge significance. To understand a critical value you have seen the sampling distribution or a theoretical distribution curve that estimates a sampling distribution. If the test statistic falls in one of the tails of the sampling distribution or in the tail of the theoretical curve, this indicates significance. If the test statistic does not fall in one of the tails of the theoretical distribution curve, the sample data does not significantly disagree with the null hypothesis. A key question is where do the tails start? That is what the critical value is calculating. The critical value is the cutoff for where the tails of the distribution begin. We just need to compare the test statistic to the critical value and see if the test statistic falls in the tail. When a computer program calculates the test statistic and critical value for you, it is always good to draw the theoretical curve and use the critical values to visualize the tails. Now ask yourself if the test statistic is in the tail or not.

Significance Rule for Test Statistics and Critical Values

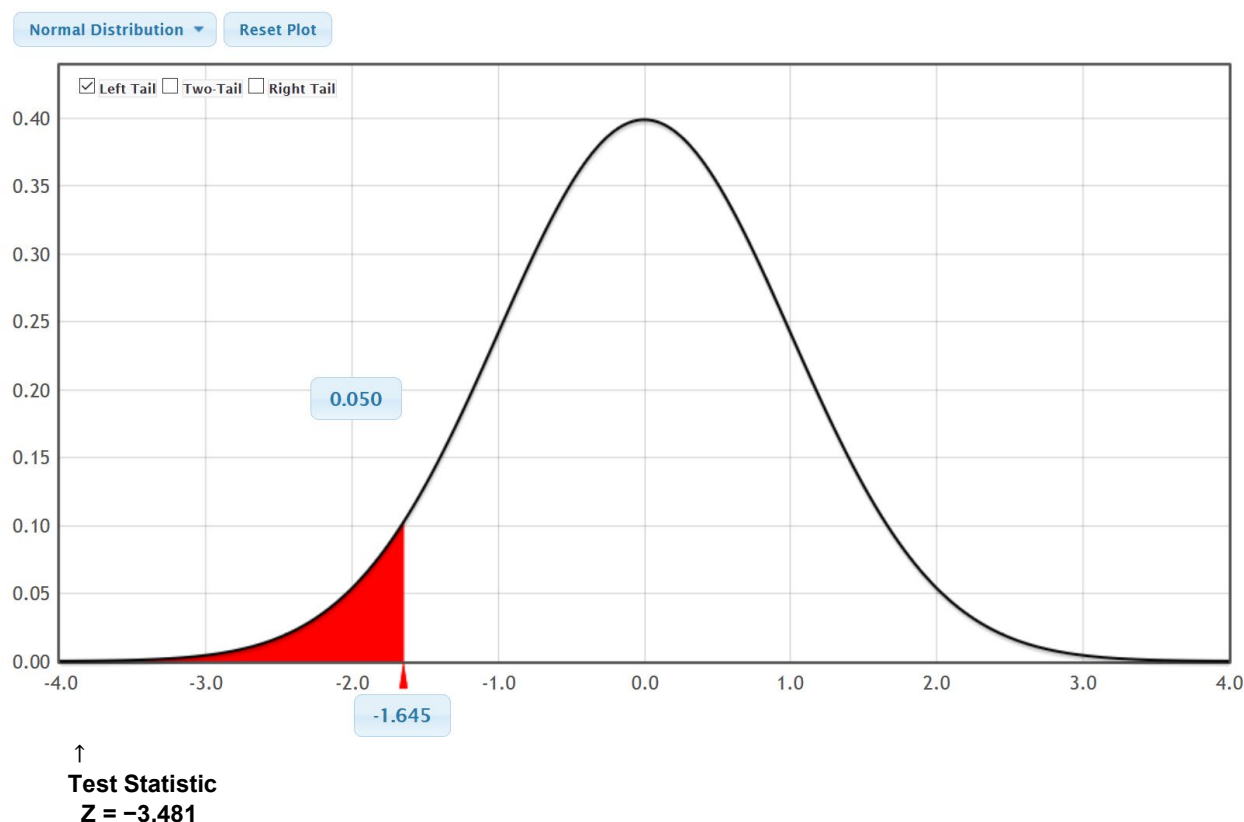
- **If the test statistic falls in the tail determined by the critical value (or values), then the sample data significantly disagrees with the null hypothesis.**
- **If the test statistic does NOT fall in the tail determined by the critical value (or values), then the sample data does NOT significantly disagree with the null hypothesis.**

Left-tailed test statistic example

Suppose we are doing a left tailed hypothesis test for proportions and the computer calculates a test statistic of $Z = -3.481$ and a critical value of -1.645 . The Z-distribution is a normal curve. So we need to see if the test statistic falls in the left tail. Negative numbers can be particularly challenging. Remember on the number line, -3.481 is less than -1.645 , so the test statistic does fall in the left tail. This means that the sample data significantly disagrees with the null hypothesis.

A nice way to visualize this principle is to go to the “Theoretical Distributions” menu in StatKey at www.lock5stat.com. Click on “normal”, then “Left Tail”. Put the critical value of -1.645 in the bottom box.



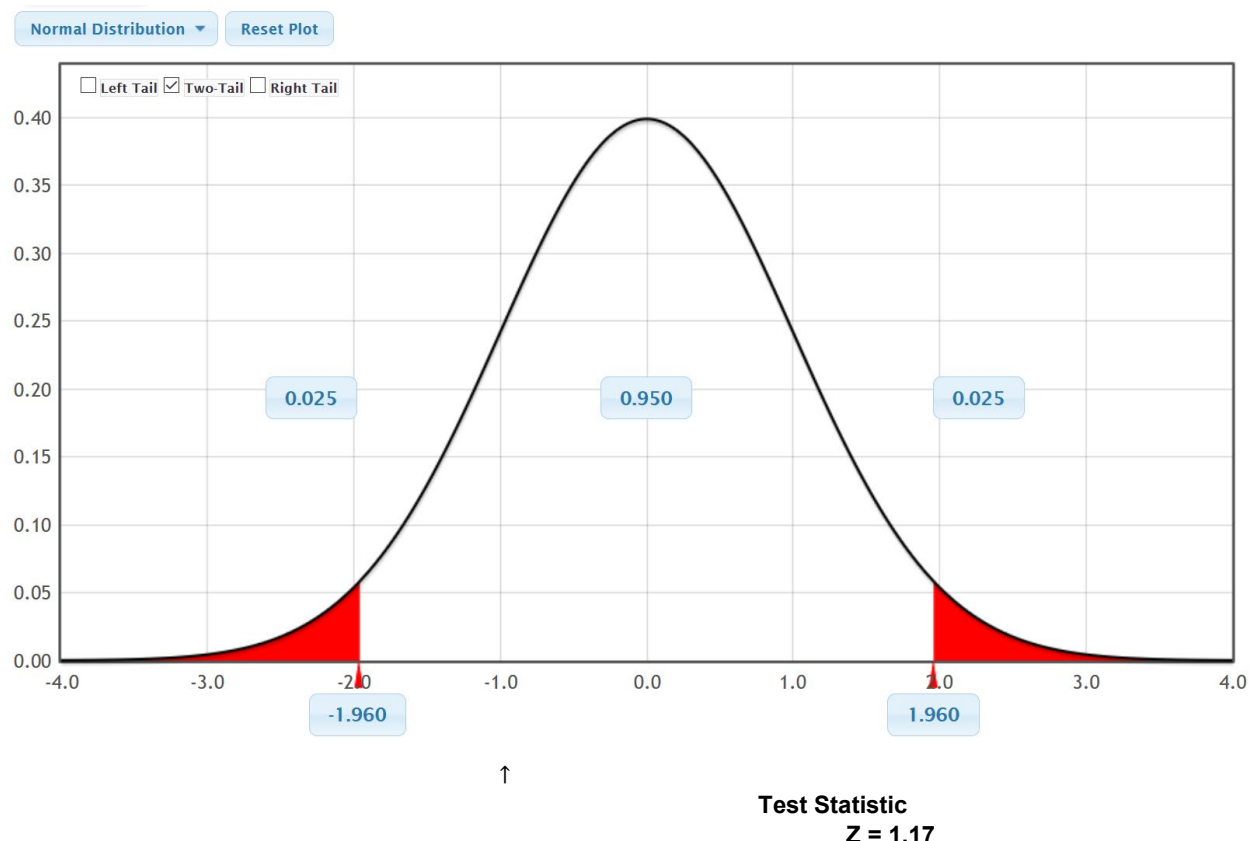


Two-tailed Test Statistic Example

Suppose we are doing a two-tailed hypothesis test for proportions. The computer calculated a Z-test statistic of $Z = +1.17$ and Critical Values of ± 1.96 .

Notice that in a two-tailed test, there are two critical values. In normal distributions like Z or T, two-tailed critical values are a certain number of standard errors above or below and so the positive and negative number of standard errors. In this example we see the “plus or minus” notation for the critical value ± 1.96 . This really means that the lower critical value for the left tail is -1.96 and the upper critical value for the right tail is $+1.96$. In non-normal distributions like F or χ^2 , the upper and lower critical values are not plus or minus. They all work the same way though. Is my test statistic in the tail. If we draw the normal Z-curve with the right tail starting at the upper critical value of $+1.96$ and left tail starting at -1.96 , we can see if the test statistic of $+1.17$ falls in the tail. Again, we can use StatKey to visualize the curve and the critical values. Go to the “Theoretical Distributions” menu in StatKey at www.lock5stat.com. Click on “normal”, then “Two Tail”. Put the critical value of -1.96 in the bottom box in the left tail and $+1.96$ in the bottom box of the right tail. Where does our test statistic fall?





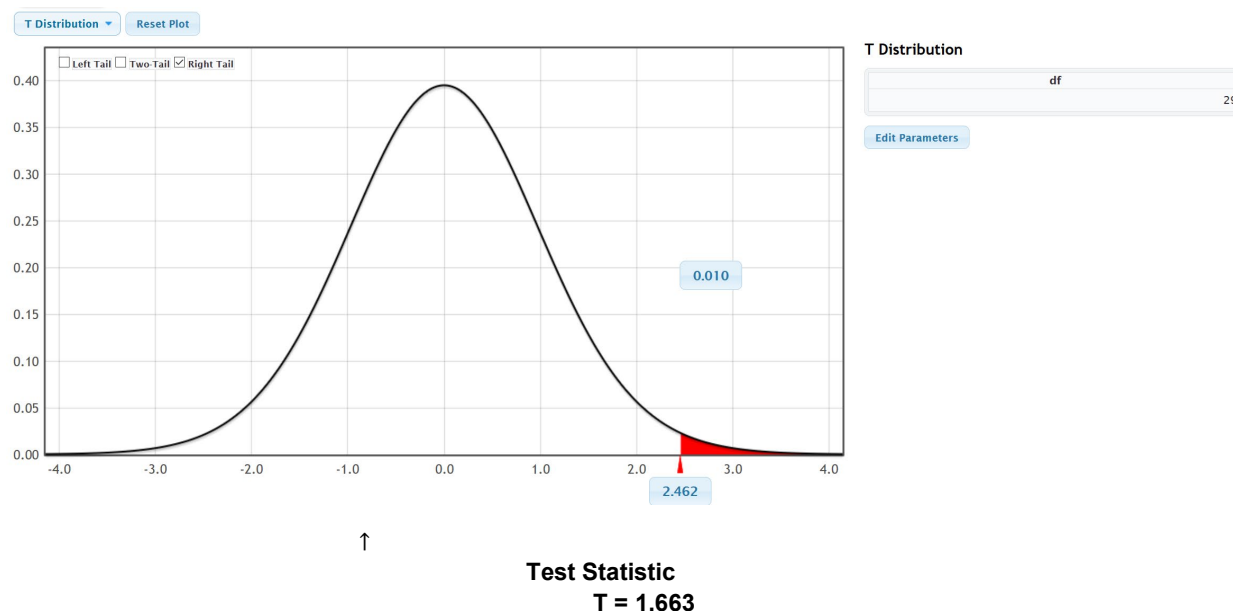
Notice our test statistic did not fall in one of the tails. So our test statistic and critical values are indicating that the sample data does not significantly disagree with the null hypothesis.

Right-tailed Test Statistic Example

Suppose we are doing a right tailed hypothesis test for means and a degrees of freedom of 29. The computer calculated a T-test statistic of $T = +1.663$ and a critical value of $+2.462$.

Like the Z-distribution, the T-distribution is also normal. So if this was a two-tailed test, we would see the “ \pm ” notation. This is a right tailed test though, so we see there is one upper critical value for the right tail. Draw the normal T-curve with the right tail starting at the upper critical value of $+2.462$. You can have StatKey draw it for you if you wish. Go to the “Theoretical Distributions” menu in StatKey at www.lock5stat.com. Click on “t”, then put in the degrees of freedom of 29. Now click “Right Tail” and put the critical value of $+2.462$ in the bottom box below the right tail. Does our test statistic fall in the tail?





Notice that our T-test statistic of 1.663 did not fall in the tail determined by the critical value. This again tells us that our sample data does not significantly disagree with the null hypothesis.

Significance Levels (Alpha Levels) (α)

The significance level (or alpha level) is an important number in a hypothesis test. It is often denoted by the Greek letter “alpha” or “ α ”. When scientists perform a hypothesis test, they choose a significance level for the test. This number is very important. It is tied to the critical value and is also very important in understanding the amount of sampling variability in a hypothesis test. Do you recall the confidence levels from the last chapter? Think of the significance level (α) as the opposite of the confidence level ($1 - \alpha$).

<u>Confidence Levels ($1 - \alpha$)</u>	<u>Significance Levels (α)</u>
90% (0.90)	10% (0.10)
95% (0.95)	5% (0.05)
99% (0.99)	1% (0.01)

If you want to be 95% confident, you will use a 5% significance level in your hypothesis test. Similarly, if you want to be 90% confident, you will use a 10% significance level. 99% confident corresponds to a 1% significance level.

Note: Remember that a 95% confidence level is the most common. Not surprisingly, a 5% significance level is also the most common. A good rule of thumb, is that if you do not know what significance level to use, use 5% ($\alpha = 0.05$).

One-population proportion Z-test statistic

The test statistic used for a one-population proportion (%) hypothesis test is the Z-test statistic. The Z-test statistic counts the number of standard errors that the sample proportion (\hat{p}) is above or below the population proportion (π) in the null hypothesis. If you recall from previous chapters, the number of standard errors or standard deviations is often called a “Z-score”. Not surprising, the one-population proportion test statistic is a Z-score.

Key question: How can we tell if the Z-test statistic indicates a significant disagreement between the sample proportion and the population proportion in the null hypothesis?



Remember, to know if a test statistic indicates a significant difference, compare the test statistic to a critical value. If the test statistic fall in the tail of the distribution determined by the critical value, then it is a significant difference.

If you remember, there are three famous critical value Z-scores for 90%, 95% and 99% confidence (± 1.645 , ± 1.96 or ± 2.576). These numbers are often used to see if our Z-test statistic is significant.

One-population proportion Z-test statistic example

Suppose we want to test the claim that the population percentage is 25%. Let's look at the percentage problem where the population proportion (π) is 0.25 (25%) and the sample proportion (\hat{p}) is 0.22 (22%). We know the sample value is % lower, but we do not know if that is significant. Another important bit of information is the sample size. In this case, it was 100.

$$H_0: \pi = 0.25 \text{ (Claim)}$$

$$H_A: \pi \neq 0.25$$

To find out if our sample data disagrees with the null hypothesis, let's calculate the test statistic.

Formulas for Test Statistics often follow general patterns. A one-population proportion or mean test statistic counts how many standard errors that the sample statistic is above or below the population parameter in the null hypothesis. Here is the general formula.

$$\frac{(\text{Sample Statistic} - \text{Population Parameter})}{\text{Standard Error}}$$

Remember in the last unit we saw that statisticians often used formulas to approximate the standard error. For one-proportion confidence intervals the standard error was $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. In a hypothesis test we have an idea of what the population proportion (π) is, so we prefer to use π instead of \hat{p} in the standard error estimation formula.

Here is the formula for a one-population proportion Z-test statistic.

$$Z = \frac{(\hat{p} - \pi)}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Let's plug in our numbers. Remember that the sample proportion $\hat{p} = 0.22$, the population proportion $\pi = 0.25$ and the sample size $n = 100$.

$$Z = \frac{(\hat{p} - \pi)}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{(0.22 - 0.25)}{\sqrt{\frac{0.25(1-0.25)}{100}}} = \frac{(0.22 - 0.25)}{\sqrt{\frac{0.25(0.75)}{100}}} \approx \frac{(-0.03)}{0.0433} \approx -0.69$$

It is always very important to explain your test statistic. In this one population proportion hypothesis test, the Z-test statistic is counting how many standard errors the sample proportion (\hat{p}) is above or below the population proportion (π). If the test statistic is negative, it is "below". If the test statistic is positive, it is "above".

Z-test statistic sentence:

The sample proportion (0.22) was 0.69 standard errors below the population proportion (0.25).

OR

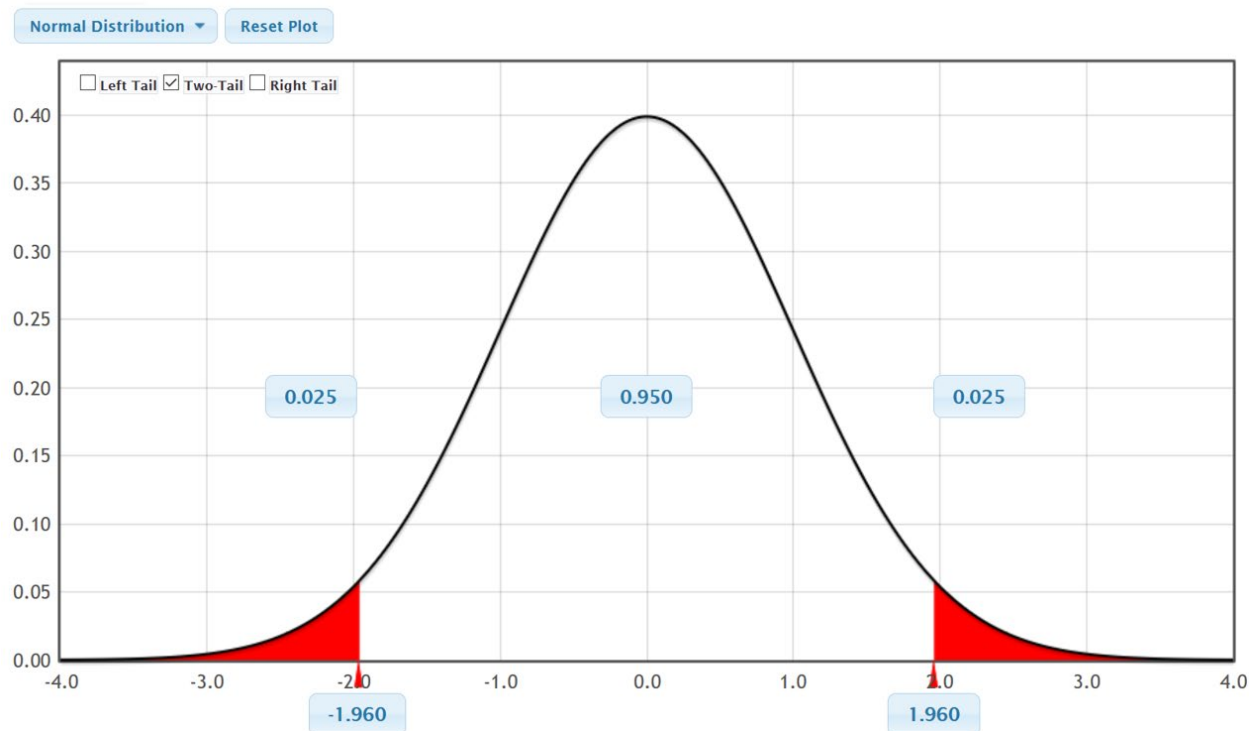
The sample percentage (22%) was 0.69 standard errors below the population percentage 25%.



Is it significant? To determine if this is significant, we should calculate the critical value. Computer programs usually calculate the test statistic and critical value for us. We also learned in the last chapter that we can use StatKey to calculate a critical value.

To calculate a critical value, you must first determine what significance level you plan to use and if your hypothesis test is left-tailed, right-tailed or two-tailed. In this example, the data scientist used a 5% significance level. Notice the alternative hypothesis was “ \neq ”, so this is a two-tailed test.

Go to www.lock5stat.com and click on “StatKey”. Under the theoretical distributions menu click on “normal”. This is the menu for calculating critical value Z-scores. Leave the mean at zero and the standard deviation as one. Click two tail. The significance level is the probability in the tail. Since there are two tails, the significance level will be broken in half with 2.5% in each tail. The numbers on the bottom are the critical values.



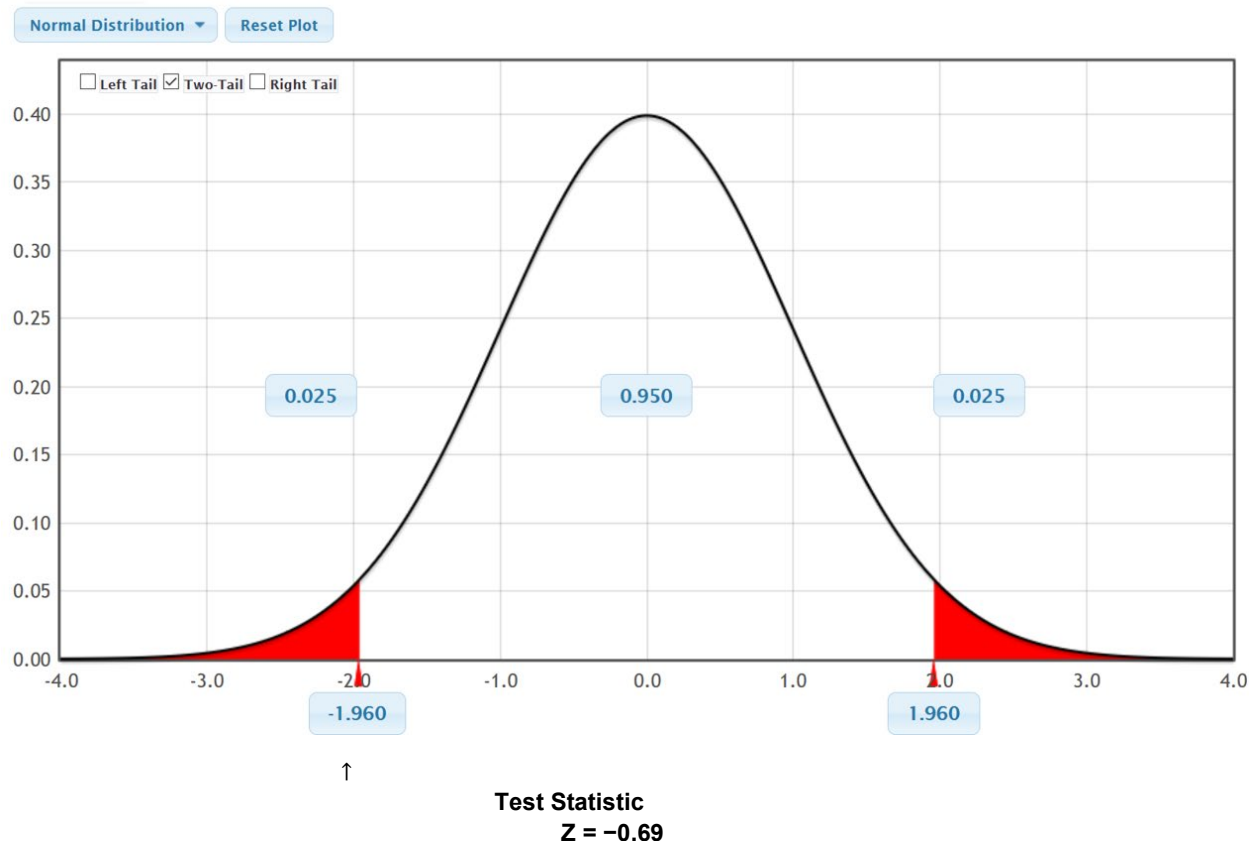
So we see that our critical values for this hypothesis test are ± 1.96 . Remember, our test statistic is $Z = -0.69$. Let's use our significance rule to determine if the sample data significantly disagreed with the null hypothesis.

Significance Rule for Test Statistics and Critical Values

- If the test statistic falls in the tail determined by the critical value (or values), then the sample data significantly disagrees with the null hypothesis.
- If the test statistic does NOT fall in the tail determined by the critical value (or values), then the sample data does NOT significantly disagree with the null hypothesis.

Did our Z-test statistic of $Z = -0.69$ fall in one of the tails?





Our test statistic did not fall in the tail, so is not significant. The sample proportion and population proportion in the null hypothesis are not significantly different. They are actually pretty close. This also tells us that the sample data does not significantly disagree with the null hypothesis.

Statcato: Most people working in statistics today do not calculate test statistics and critical values. Statistics computer programs can calculate them for us. In Statcato, go to the statistics menu, then click on hypothesis test. Click on "one-population proportion". Most computer programs require the number of success (events) to calculate proportions. In the last example, they did not tell us the number of successes. They just told us the total sample size (100) and the sample proportion (0.22). If you multiply them, you get the number of successes or events.

Number of Successes (events) = $100 \times 0.22 = 22$

So in the one-population proportion hypothesis test menu in Statcato, go to the summarized data and enter 22 for the number of events and 100 for the total number of trials. Put in not equal for the alternative hypothesis, 0.25 for the hypothesized proportion, and a significance level of 0.05. Here is the Statcato printout. Notice the test statistic is close to our calculation and the critical values are the same as what we calculated with StatKey.



Hypothesis Test: 1-Population Proportion

Help F1

Inputs

☐ Samples in column:

☒ Summarized sample data:

Number of events: 22

Number of trials: 100

Significance

☒ Significance Level: 0.05 0 - 1.00 (e.g. 0.05)

☐ Confidence Level: 0.95 0 - 1.00 (e.g. 0.95)

Alternative Hypothesis

Alternative Hypothesis: Not Equal to

Hypothesized Proportion: 0.25

OK Cancel

Hypothesis Test - One population proportion: confidence level = 0.95

Input: Summary data

Null hypothesis: $p = 0.25$

Alternative hypothesis: $p \neq 0.25$

N	Sample Proportion	Significance Level	Critical Value	Test Statistic Z	p-Value
100	0.22	0.05	-1.96, 1.96	-0.693	0.4884

Note about Sample Size

Sample size plays a key role in significance. Let's look at the previous example but with a sample size of 1000.

$$Z = \frac{(\hat{p} - \pi)}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{(0.22 - 0.25)}{\sqrt{\frac{0.25(1-0.25)}{1000}}} = \frac{(0.22 - 0.25)}{\sqrt{\frac{0.25(0.75)}{1000}}} = \frac{(-0.03)}{0.0137} \approx -2.19$$

If our sample size had been 1000, then the sample percentage of 22% is significantly lower than our population percentage of 25%. In fact, our sample percentage of 22% is 2.19 standard errors below the population percentage of 25%. A test statistic of $Z = -2.19$ would have fallen in the left tail since it is less than the lower critical value of -1.96 . At a sample size of 1000, 22% is significantly less than 25%.

One-population mean T-test statistic

The test statistic used for a one-population mean hypothesis test is the T-test statistic. Like the Z-test statistic, the one-population T-test statistic also counts the number of standard errors that the sample statistic is from the population parameter in the null hypothesis. In this case, it will count the number of standard errors that the sample mean (\bar{x}) is above or below the population mean (μ) in the null hypothesis.

Key question: How can we tell if the T-test statistic indicates a significant disagreement between the sample mean (\bar{x}) and the population mean (μ) in the null hypothesis?



Remember, to know if a test statistic indicates a significant difference, compare the test statistic to a critical values. For a significant difference, the test statistic should fall in the tail of the distribution determined by the critical values.

In our last chapter, we saw that T-score critical values are organized by degrees of freedom. For a one-population hypothesis test, the degrees of freedom is the sample size – 1 or “n – 1”.

One-population mean T-test statistic example

An article in a health magazine claims that the mean average weight of all men is less than 180 pounds. A random sample of 40 men found that the sample mean was 172.55 pounds with a sample standard deviation of 26.327 pounds. Use a 10% significance level and the random sample data to test this claim.

$$H_0 : \mu = 180$$

$$H_A : \mu < 180 \text{ (Claim)}$$

Notice that the sample mean of 172.55 pounds is 7.45 pounds less than the population mean of 180 pounds. Is that significant?

The answer again is we don't know yet. We would need calculate a test statistic and a critical value to see if 7.45 pounds is a lot in this situation. Notice that since the alternative hypothesis is less than, this is a left-tailed test.

Here is the formula for calculating test statistics to compare sample and population means. Notice it follows the same general pattern and seeks to count how many standard errors the sample statistic is above or below the population parameter in the null hypothesis. In this case, the one-population T-test statistic calculates the number of standard errors that the sample mean (\bar{x}) is above or below the population mean (μ) in the null hypothesis. If the test statistic comes out negative it will be “below” and if the test statistic is positive, it will be “above”. Notice also that we are using the same standard error estimation formula that we used for confidence intervals.

$$T = \frac{(\text{Sample Mean} - \text{Population Mean})}{\text{Standard Error}} = \frac{(\bar{x} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)}$$

To calculate the test statistic, plug in the sample mean $\bar{x} = 172.55$, the population mean $\mu = 180$, the sample standard deviation $s = 26.327$, and sample size $n = 40$.

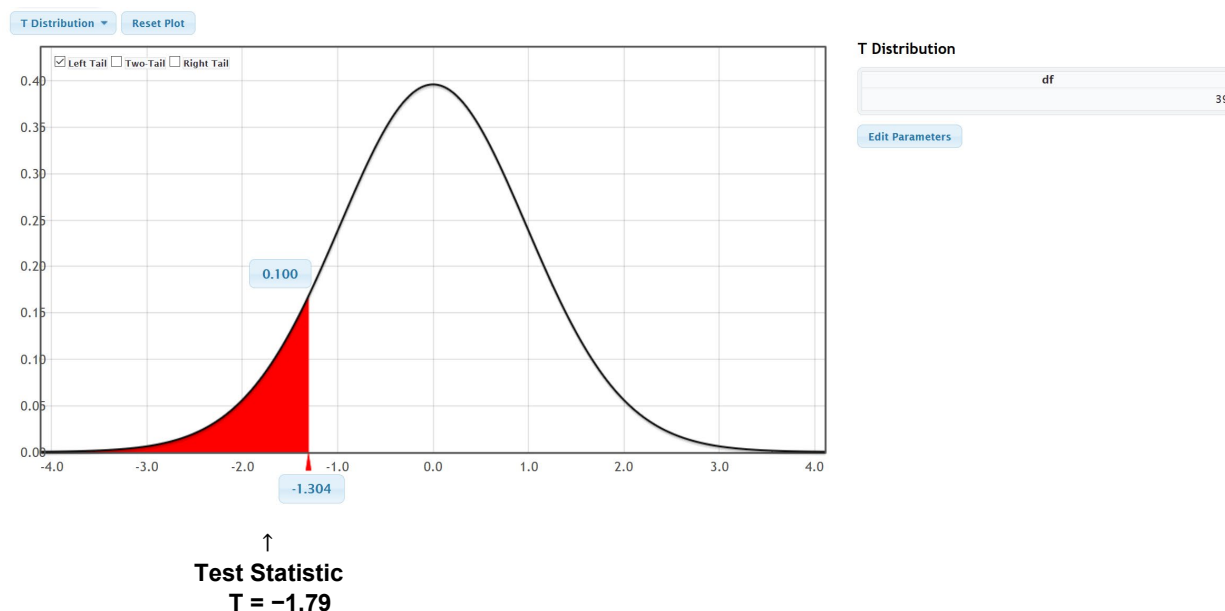
$$T = \frac{(\bar{x} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{(172.55 - 180)}{\left(\frac{26.327}{\sqrt{40}}\right)} \approx \frac{(-7.45)}{4.1627} \approx -1.79$$

Test-statistic sentence: The sample mean of 172.55 pounds is 1.79 standard error below the population mean of 180 pounds.

Is it significant?

Again, we will use StatKey to calculate the critical value. Go to the “Theoretical Distributions” menu and click on “T”. In this problem the degrees of freedom is $40 - 1 = 39$. We also used a 10% significance level for this left-tailed test. After entering the degrees of freedom into StatKey, click on left tail, and put in 10% (0.10) for the tail proportion. Does the test statistic of $T = -1.79$ fall in the tail?





Notice that since the test statistic $T = -1.79$ is less than the critical value of -1.304 , the T-test statistic does fall in the left tail. So our sample mean of 172.55 pounds is significantly lower than the population mean of 180 pounds and the sample data significantly disagrees with the null hypothesis.

Statcato: Again, we can calculate the test statistic and critical value with a program like Statcato. Go to the “statistics” menu again and click on “hypothesis tests”. Now click on “one-population mean”. Under summarized data, put in the sample size of 40, sample mean of 172.55, and the sample standard deviation of 26.327. Change the significance level to 0.10, the alternative hypothesis to less than, and the hypothesized mean as 180. Then push ok. The printout is given below. Notice the test statistic is the same as our calculation and the critical value is the same as what we got with StatKey.

Hypothesis Test: 1-Population Mean

Help F1

Inputs

☐ Samples in column: ☒ Summarized sample data:

Size: 40

Mean: 172.55

Standard deviation: 26.327

Population Standard Deviation

Population standard deviation:

☐ Known: - use z distribution

☒ Unknown - use t distribution

Significance

☒ Significance Level: 0.10 0 - 1.00 (e.g. 0.05)

☐ Confidence Level: 0.90 0 - 1.00 (e.g. 0.95)

Alternative Hypothesis

Alternative Hypothesis: Less than

Hypothesized Mean: 180

OK Cancel



Hypothesis Test - One Population Mean: confidence level = 0.90

Input: Summary data

 σ unknown (using t distribution)Null hypothesis: $\mu = 180.0$ Alternative hypothesis: $\mu < 180.0$

N	Sample Mean	Stdev s	Significance Level	Critical Value	Test Statistic	p-Value
40	172.55	26.327	0.10	-1.304	-1.790	0.0406

One-population variance χ^2 -test statistic

When doing a hypothesis test about a population standard deviation (σ), we often prefer to use the population variance (σ^2). Remember, the variance is the square of the standard deviation. You can also think of the standard deviation as the square root of the variance. In a one population variance or standard deviation hypothesis test, we will compare the sample variance (s^2) to the population variance (σ^2).

One-population variance χ^2 -test statistic Example

While many people believe that the population standard deviation for men's weight is 20 pounds. This also implies that the population variance for men's weight is 400 square pounds. A random sample of 40 men has a sample standard deviation of 26.327 pounds and a sample variance is 693.111 square pounds. Let's use a 1% significance level for this problem. Notice that the test can be done using the standard deviation or the variance.

$$H_0 : \sigma = 20 \ (\sigma^2 = 400) \text{ (Claim)}$$

$$H_A : \sigma \neq 20 \ (\sigma^2 \neq 400)$$

Notice that the sample standard deviation of 26.327 pounds is 6.327 pounds more than the population standard deviation of 20 pounds.

Is that significant?

To determine this, we will need a test statistic. In our discussions about sampling distributions in the last chapter, we learned that the sampling distribution for variance is rarely normal and usually skewed to the right. For this reason we cannot use a Z or T-test statistic. These are based on normal sampling distributions. We saw in the last chapter that for one-population variance and standard deviation, we use the Chi-Squared distribution (χ^2) with a degrees of freedom $n - 1$.

Chi-Squared One-population Variance Test Statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

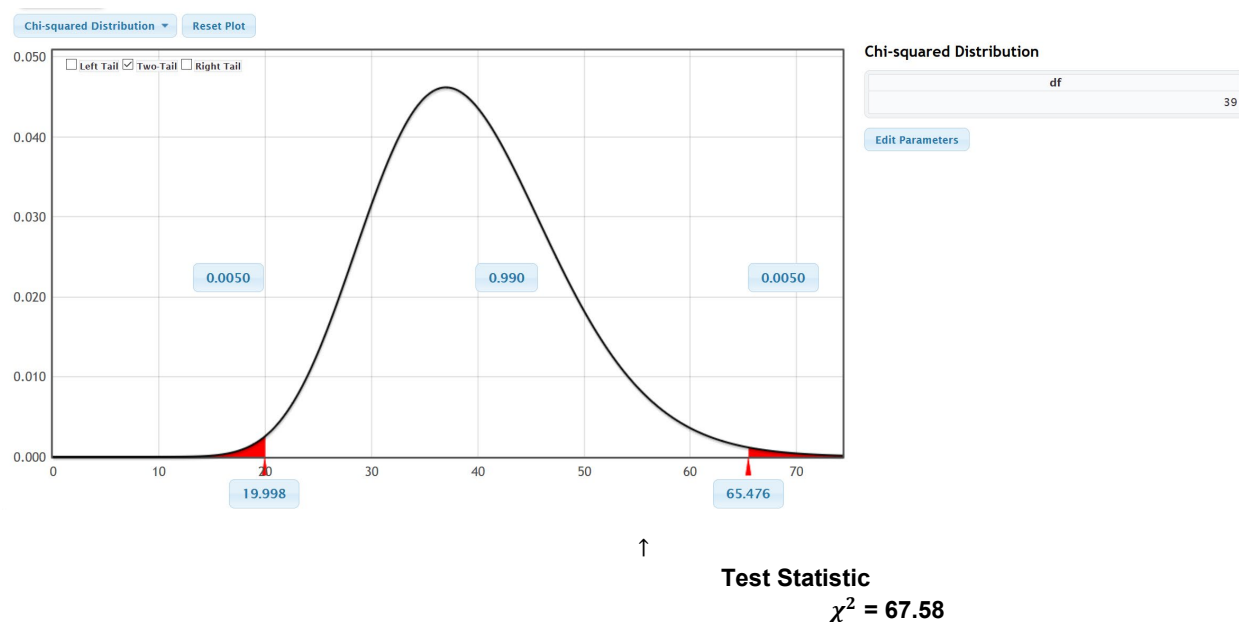
Let's calculate the Chi-Squared test statistic for this problem. The degrees of freedom for this problem is $df = n - 1 = 40 - 1 = 39$. The sample standard deviation is $s = 26.327$, so the sample variance is $s^2 = 693.111$. The population standard deviation in the null hypothesis is 20 pounds, so the population variance is $\sigma^2 = 400$.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(40-1)26.327^2}{20^2} = \frac{(39)693.11}{400} \approx 67.58$$

Is this significant? Again, we will need to calculate the critical values. Let's go to StatKey at www.lock5stat.com. Under the theoretical distributions menu click on " χ^2 ". Put in the degrees of freedom 39 and click "Two Tail".

Since this is a two-tailed test we will need to divide the significance level of 1% in half and put 0.5% (0.005) in each tail. What are the critical values? Does the test statistic fall in the tail?





First notice that the Chi-squared distribution is not perfectly normal. The F-distribution is very similar. Also the upper and lower critical values are not positive or negative like Z or T. It still works the same way though. We see that we have divided the significance level 0.01 in half and put 0.005 in each tail. The upper critical value is 65.476 and lower critical value is 19.998.

We also see that the χ^2 test statistic of 67.58 is higher than the upper critical value of 65.476, so falls in the right tail. The sample standard deviation significantly disagrees with the population standard deviation and the sample data significantly disagrees with the null hypothesis.

Statcato: Remember, it is always preferable to use technology to calculate test statistics. To calculate the χ^2 -test statistic and critical value with Statcato, go to the "statistics" menu, click on "hypothesis tests", then click on "one-population variance". Put in the sample size of 40 and either the sample standard deviation (26.327 pounds) or the sample variance. Change the significance to 1% (0.01). We also need to put in the population standard deviation (20) or the population variance and change the alternative hypothesis to "Not Equal to". Now press "OK".

Hypothesis Test: 1-Population Variance

Help F1

Inputs

☐ Samples in column:

☒ Summarized sample data:

Sample Size: 40

☐ Variance:

☒ Standard deviation: 26.327

Significance

☒ Significance Level: 0.01 0 - 1.00 (e.g. 0.05)

☐ Confidence Level: 0.99 0 - 1.00 (e.g. 0.95)

Alternative Hypothesis

Alternative Hypothesis: Not Equal to

☐ Hypothesized variance:

☒ Hypothesized standard deviation: 20

OK Cancel



Hypothesis Test - One population variance: confidence level = 0.99

Input: Summary data

Null hypothesis: $\sigma^2 = 400.0$ Alternative hypothesis: $\sigma^2 \neq 400.0$

N	Sample Stdev s	Sample Var s^2	Significance Level	Critical Value	Test Statistic	p-Value
40	26.327	693.111	0.01	19.996, 65.476	67.578	0.0061

Notice the test statistic is close to what we calculated above and the critical values are very close to our StatKey calculation.

Important Notes:

- *It is important to understand how test statistics work and what the formulas represent. It is not important to calculate these by hand with a calculator. Statistics programs can calculate the test statistic quickly and with much better accuracy. It is important that you can explain the test statistic and what it tells us about significance.*
- *Whether or not a test statistic is significant can be difficult to interpret. Computer programs often give critical values to compare the test statistic to so that you can know if it was significant. It is important to draw a picture and visualize where the critical values are and where the tails begin. Here is the significance rule for test statistics and critical values.*

Significance Rule for Test Statistics and Critical Values

- **If the test statistic falls in the tail determined by the critical value (or values), then the sample data significantly disagrees with the null hypothesis.**
 - **If the test statistic does NOT fall in the tail determined by the critical value (or values), then the sample data does NOT significantly disagree with the null hypothesis.**
-



Practice Problems Section 3B

(#1-20) For each of the following, use the given test statistic and critical value or values to answer the following questions.

- Draw the indicated distribution and use the critical values to label the tails.
- Does the test statistic fall in one of the tails or not?
- Does the sample data significantly disagree with the null hypothesis? Explain how you know.

	Tail	Test Statistic	Critical Value
1	Two	$Z = 2.47$	± 1.96
2	Left	$T = -3.318$	-1.747
3	Right	$\chi^2 = 6.943$	12.33
4	Right	$F = 1.126$	3.881
5	Left	$Z = -1.33$	-1.645
6	Two	$T = 1.994$	± 2.738
7	Right	$\chi^2 = 18.441$	6.972
8	Right	$F = 7.509$	3.469
9	Two	$Z = -2.72$	± 2.576
10	Left	$T = -3.871$	-2.114
11	Left	$Z = -1.884$	-2.576
12	Right	$T = 0.472$	1.577
13	Two	$\chi^2 = 11.943$	2.346 & 9.841
14	Right	$F = 5.218$	2.791
15	Left	$Z = -2.712$	-1.96
16	Two	$T = 1.138$	± 2.005
17	Right	$\chi^2 = 38.644$	12.359
18	Right	$F = 1.528$	2.467
19	Left	$Z = -0.72$	-2.576
20	Two	$T = -2.871$	± 2.334

(#21-23) Use the “theoretical distributions” menu in StatKey at www.lock5stat.com to look up the following critical values. Click on the button that says “normal”. Then answer the questions.

21. Z-test statistic = 2.36
Two-tailed test
Significance Level = 5% (0.025 in each tail)

Critical Values =

Does the sample data significantly disagree with the null hypothesis? Explain why.

22. Z-test statistic = -1.48
Left-tailed test
Significance Level = 1% (0.01 in left tail)

Critical Value =

Does the sample data significantly disagree with the null hypothesis? Explain why.

23. Z-test statistic = 2.02
Right-tailed test
Significance Level = 10% (0.10 in right tail)

Critical Value =

Does the sample data significantly disagree with the null hypothesis? Explain why.



(#24-26) Use the “theoretical distributions” menu in StatKey at www.lock5stat.com to look up the following critical values. Click on the button that says “t”. Then answer the questions.

24. T-test statistic = -1.773

Two-tailed test

Degrees of Freedom = 29

Significance Level = 1% (0.005 in each tail)

Critical Values =

Does the sample data significantly disagree with the null hypothesis? Explain why.

25. T-test statistic = 2.871

Right-tailed test

Degrees of Freedom = 34

Significance Level = 10% (0.10 in right tail)

Critical Value =

Does the sample data significantly disagree with the null hypothesis? Explain why.

26. T-test statistic = -1.144

Left-tailed test

Degrees of Freedom = 49

Significance Level = 5% (0.05 in left tail)

Critical Value =

Does the sample data significantly disagree with the null hypothesis? Explain why.

(#27-29) Use the “theoretical distributions” menu in StatKey at www.lock5stat.com to look up the following critical values. Click on the button that says “ χ^2 ”. Then answer the questions.

27. χ^2 -test statistic = 38.725

Right-tailed test

Degrees of Freedom = 29

Significance Level = 5% (0.05 in right tail)

Critical Value =

Does the sample data significantly disagree with the null hypothesis? Explain why.

28. χ^2 -test statistic = 15.846

left-tailed test

Degrees of Freedom = 39

Significance Level = 10% (0.10 in left tail)

Critical Value =

Does the sample data significantly disagree with the null hypothesis? Explain why.

29. χ^2 -test statistic = 5.119

two-tailed test

Degrees of Freedom = 19

Significance Level = 1% (0.005 in each tail)

Critical Value =

Does the sample data significantly disagree with the null hypothesis? Explain why.



(#30-32) Use the following one-population test statistic formula to calculate the one-population proportion Z-test statistic. Then write a sentence to explain the test statistic.

$$\text{One-Population Proportion Z-Test Statistic} = \frac{(\text{Sample Proportion} - \text{Population Proportion})}{\text{Standard Error}}$$

30. Sample Proportion (\hat{p}) = 0.317
Population Proportion (π) = 0.25
Standard Error = 0.031

Z-test statistic =

Test Statistic Sentence:

31. Sample Proportion (\hat{p}) = 0.835
Population Proportion (π) = 0.9
Standard Error = 0.053

Z-test statistic =

Test Statistic Sentence:

32. Sample Proportion (\hat{p}) = 0.112
Population Proportion (π) = 0.2
Standard Error = 0.047

Z-test statistic =

Test Statistic Sentence:

(#33-35) Use the following one-population test statistic formula to calculate the one-population mean T-test statistic. Then write a sentence to explain the test statistic.

$$\text{One-Population Mean T-Test Statistic} = \frac{(\text{Sample Mean} - \text{Population Mean})}{\text{Standard Error}}$$

33. Sample Mean (\bar{x}) = 135.7 mg
Population Mean (μ) = 100 mg
Standard Error = 23.9 mg

T-test statistic =

Test Statistic Sentence:

34. Sample Mean (\bar{x}) = 89.26 °F
Population Mean (μ) = 89.6 °F
Standard Error = 0.108 °F

T-test statistic =

Test Statistic Sentence:

35. Sample Mean (\bar{x}) = 52.71 thousand dollars
Population Mean (μ) = 60 thousand dollars
Standard Error = 6.42 thousand dollars

T-test statistic =

Test Statistic Sentence:



Section 3C – P-value and Significance Levels

Vocabulary

Population: The collection of all people or objects to be studied.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about “no change”, “no relationship” or “no effect”.

Sampling Variability: Also called “random chance”. The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true. If the P-value is close to zero (lower than the significance level) is unlikely to have happened because of sampling variability. If the P-value is too large (higher than the significance level), then the sample could have occurred because of sampling variability.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and if sampling variability is likely to be involved in the hypothesis test.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value and standard error without a formula.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

Critical Value: We compare a test statistic to this number to determine if the sample data significantly disagrees with the null hypothesis. If the absolute value of the test statistic is higher than the absolute value of the critical value, then the sample data significantly disagrees with the null hypothesis.

Introduction: There is a dilemma in hypothesis testing. In a hypothesis test, we want to determine if the sample data disagrees with the null hypothesis, but there is a problem. The principle of sampling variability (random chance) tells us that random samples will almost always be different than population parameters in the null hypothesis. So even if the population parameter in the null hypothesis is correct, my random sample data will still disagree with it. So how can we use random sample data to ever decide about the accuracy of a population parameter? Random samples almost always disagree with the null hypothesis.



The real question is why does the random sample data disagree? There are only two possible answers to that question and these are at the heart of the problem.

1. The random sample disagrees because the null hypothesis is wrong.

OR

2. The null hypothesis is correct and the random sample data disagrees because of sampling variability (random chance).

How do we know which option is correct in a situation? Does my random sample data disagree because all random samples disagree, or does my random sample data disagree because the null hypothesis is wrong?

To answer this question, statisticians invented the P-value.

P-value

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

We see from the definition, we see several important ideas addressed.

- The P-value is a conditional probability based on the null hypothesis being true. The P-value can only be calculated by assuming the null hypothesis was true.
- The P-value is a probability that our random sample data occurs. If the null hypothesis really is correct, then what is the probability of our random sample data occurring by random chance?
- The P-value helps us understand why our random sample data disagrees with the null hypothesis. Does it disagree because of sampling variability (random chance) or not? If it is not sampling variability, then the only other alternative is that the null hypothesis is wrong.
- The P-value not only takes into account the probability of the sample data occurring, but also any other samples that disagrees even more with the null hypothesis than our random sample data. This is what is meant by “or more extreme”.

Reading your P-value

P-value can help us with the dilemma discussed above, but only if you know how to interpret it correctly. Remember, the P-value is the probability of your random sample data occurring because of sampling variability (by random chance). Does my random sample data disagree with the null hypothesis just because of sampling variability? If so, then the population parameter in the null hypothesis might be correct. If my sample data does not disagree because of sampling variability, then the only other alternative is that the null hypothesis must be wrong. In that case we will say that we “reject the null hypothesis”.

Low P-value

Scientists like the P-value to be very close to zero. The lower the P-value, the better. Remember, the P-value is measuring the probability that the random sample data or more extreme occurred because of sampling variability. If the P-value is zero (or really close to zero), then the data probably did not occur because of sampling variability.

Think of sampling variability as a confounding variable that we need to control or at least make sure it is unlikely to be the reason the sample data disagrees. If the P-value is zero, then it is unlikely to be sampling variability (random chance).



Suppose your P-value is 0.013 (1.3%). If your car has only a 1.3% probability of starting, do you think your car will start or is it unlikely to start? If your car only has a 1.3% chance of starting, it is unlikely to start. That is a good way to think about P-value. If there is only a 1.3% probability of our random sample data disagreeing by random chance, it is probably not random chance! It is unlikely to be sampling variability.

Remember our dilemma about the two options in a hypothesis test.

1. The random sample disagrees because the null hypothesis is wrong.

OR

2. The null hypothesis is correct and the random sample data disagrees because of sampling variability (random chance).

Low P-value Key Idea: If the P-value is really close to zero, it is ruling out option 2. At least we can say that it is very unlikely to be sampling variability (option 2). In that case, the only other alternative is option 1. The random sample data disagrees with the null hypothesis because those population parameters in the null hypothesis are wrong. In that case, we can reject the null hypothesis.

P-value close to zero → Unlikely to be sampling variability → Reject H_0

High P-value

Remember, the goal is to totally rule out sampling variability as the reason our random sample data disagrees. We need the P-value to be zero or at least as close to zero as possible. It doesn't take much for a P-value to be too high. For example, suppose our P-value was 0.15 (15%). Don't events with a 15% probability sometimes happen? While 15% may be a low probability, is it really low enough to totally rule out that the event will not happen? For this reason, we need the P-value to be extremely low and extremely close to zero.

So how can we know if the P-value is too high?

The answer to this is to compare the P-value to the significance level.

In the last section, we saw that scientists pick a significance level at the beginning of the hypothesis test. We also saw that the significance level is connected to the critical value and determining if the test statistic falls in the tail. The proportion in the tail is the significance level. The significance level is also called the alpha level (α) and can be thought of as the complement of the confidence levels ($1 - \alpha$). We saw in the last section, that the most common significance level chosen is 5% ($\alpha = 0.05$).

<u>Confidence Levels ($1 - \alpha$)</u>	<u>Significance Levels (α)</u>
90% (0.90)	10% (0.10)
95% (0.95)	5% (0.05)
99% (0.99)	1% (0.01)

So the P-value must be less than or equal to the significance level, to rule out sampling variability (or at least to ensure it is very unlikely to be sampling variability). If the P-value is higher than the significance level, then the sample data could have occurred because of sampling variability.

P-value less than or equal to the significance level → Unlikely to be sampling variability → Reject H_0

P-value higher than the significance level → Could be sampling variability → Fail to reject H_0

Let's talk about these rules some. Let's look again at the P-value of 15% (0.15). If we are using a 5% significance level then the rule would indicate that the random sample data could have occurred because of sampling variability. This implies that the null hypothesis could be correct, and my sample data might disagree because all samples



disagree. Does this guarantee that the null hypothesis is correct? Absolutely not. Let's go back to the car starting analogy. If my car only has a 15% probability of starting, is it guaranteed to start? No. It still has a low probability of starting, but it might start. That is the point. A high P-value does not tell us that the null hypothesis is correct for sure. It tells us that it might be correct.

So what about our dilemma? What does a high P-value tell us about our two options?

1. The random sample disagrees because the null hypothesis is wrong.

OR

2. The null hypothesis is correct and the random sample data disagrees because of sampling variability (random chance).

High P-value Key Idea: If the P-value is too high, then we cannot rule out option 2 (sampling variability). The high P-value does not guarantee it is sampling variability though. It just might be. When the P-value is large, we will not be able to tell which option is correct. The null hypothesis might be wrong. Our sample data disagrees with it after all. On the other hand, the null hypothesis might be correct and our sample data disagrees because of sampling variability. In a sense, we cannot tell which option is correct. That is why we say "Fail to reject the null hypothesis". This means that we do not have a low enough P-value to rule out sampling variability, so we cannot say for sure that the null hypothesis is wrong. It might be correct.

P-value less than or equal to the significance level → Unlikely to be sampling variability → Reject H_0

P-value higher than the significance level → Could be sampling variability → Fail to reject H_0

For this reason, high P-values are generally not preferred by data scientists. A low P-value rules out sampling variability (rules out random chance) and allows us to reject the null hypothesis and support the alternative hypothesis. A low P-value is also considered evidence. Scientific reports often require a low P-value as evidence to support their findings. When a scientist gets a high P-value, they do not have evidence. Sampling variability is involved and they cannot really say anything definitively. This does not mean that a high P-value has no value. A low P-value gives us evidence that the alternative hypothesis is probably correct. A high P-value indicates that the null hypothesis might be correct, but we do not have evidence.

Low P-value (Less than or equal to the significance level)

- Unlikely to be sampling variability
- Reject H_0
- H_A is probably correct
- We have significant evidence.

High P-value (Higher than the significance level)

- Could be sampling variability
- Fail to reject H_0
- H_0 is probably correct
- We do not have evidence.

Example 1 (Interpreting P-values)

Suppose we have a 5% significance level and a P-value = 0.0278. Convert the P-value into a percentage and write a sentence to explain the P-value. Compare the P-value to the significance level. Is this a low P-value or a high P-value. Could this be sampling variability or is it unlikely to be sampling variability? Explain your answer. Does the sample data significantly disagree with the null hypothesis or not? Explain your answer. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

P-value = 0.0278 = 2.78%



P-value Sentence: If the null hypothesis is true, there is a 2.78% probability of getting the sample data or more extreme by random chance (because of sampling variability).

P-value (2.78%) is lower than our significance level (5%), so this is a low P-value close to zero.

Since the P-value is very low (2.78%), it is unlikely to be sampling variability (unlikely to be random chance).

A low P-value means the sample data fell in the tail and has a large test statistic. That means that the sample data does significantly disagree with the null hypothesis.

Reject the null hypothesis, since the P-value is lower than the significance level and is unlikely to be sampling variability.

Example 2 (Interpreting P-values)

Suppose we have a 10% significance level and a P-value = 0.414. Convert the P-value into a percentage and write a sentence to explain the P-value. Compare the P-value to the significance level. Is this a low P-value or a high P-value. Could this be sampling variability or is it unlikely to be sampling variability? Explain your answer. Does the sample data significantly disagree with the null hypothesis or not? Explain your answer. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

P-value = $0.414 = 41.4\%$

P-value Sentence: If the null hypothesis is true, there is a 41.4% probability of getting the sample data or more extreme by random chance (because of sampling variability).

P-value (41.4%) is higher than our significance level (10%), so this is a high P-value.

Since the P-value is very high (41.4%), the sample data could have occurred because of sampling variability (could be random chance).

A high P-value means the sample data did not fall in the tail and has a small test statistic. That means that the sample data does NOT significantly disagree with the null hypothesis.

Fail to reject the null hypothesis, since the P-value is higher than the significance level and could be sampling variability.

Calculating P-values

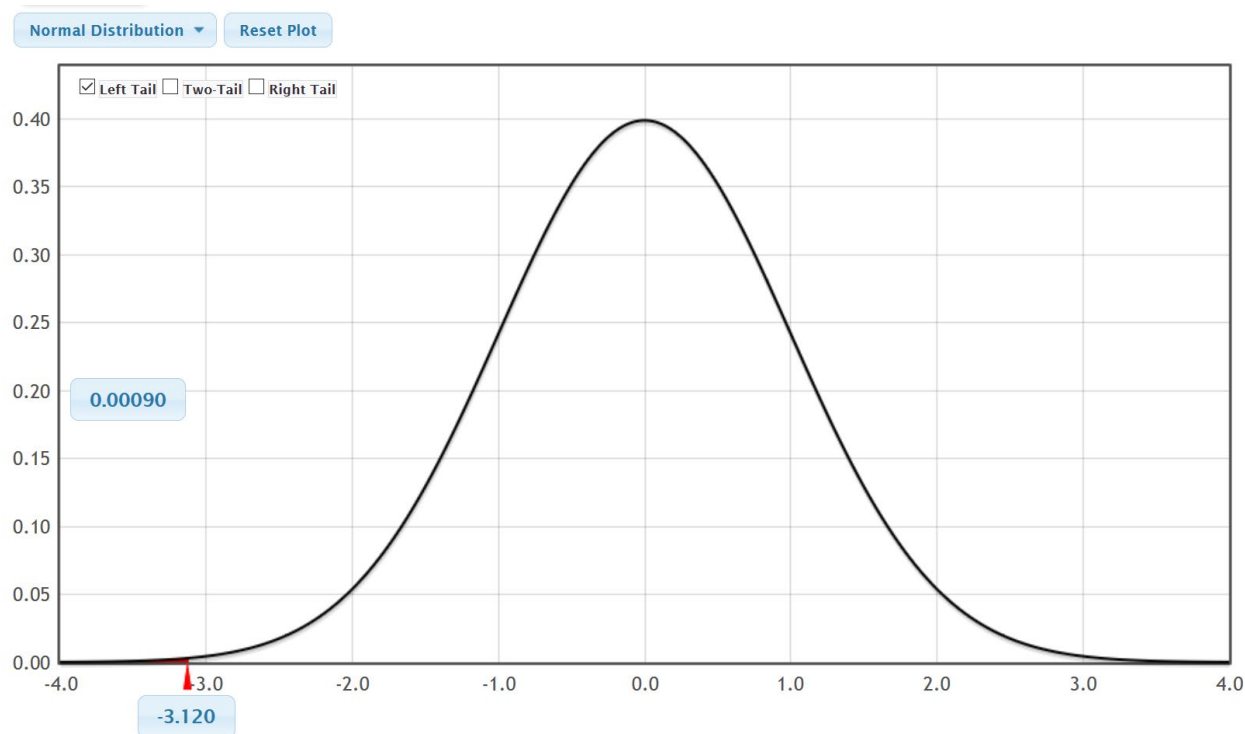
Method 1: Traditional Approach (Using the test statistic and a theoretical curve.)

One way to calculate P-value is with the test statistic and the theoretical curve that represents the sampling distribution. Remember, the P-value is the probability of getting the sample data or more extreme if the null hypothesis is true. Think of the test statistic as representing the sample data if the null hypothesis was true. "The probability of getting the sample data or more extreme" would be the proportion in the tail or tails using the test statistic as your cutoff and taking into account the type of test you are doing.

Example 1: Suppose we are doing a left tailed hypothesis test that uses the Z-test statistic for proportions. Our test statistic compares the sample data to the null hypothesis. In this example, our Z-test statistic was calculated to be $Z = -3.12$. What would the estimated P-value be?

Go to the "theoretical distributions" menu in StatKey at www.lock5stat.com and click on "normal". Click on "Left Tail". In the bottom box in the left tail, put in the test statistic of -3.12 . The proportion calculated above is the estimated P-value.



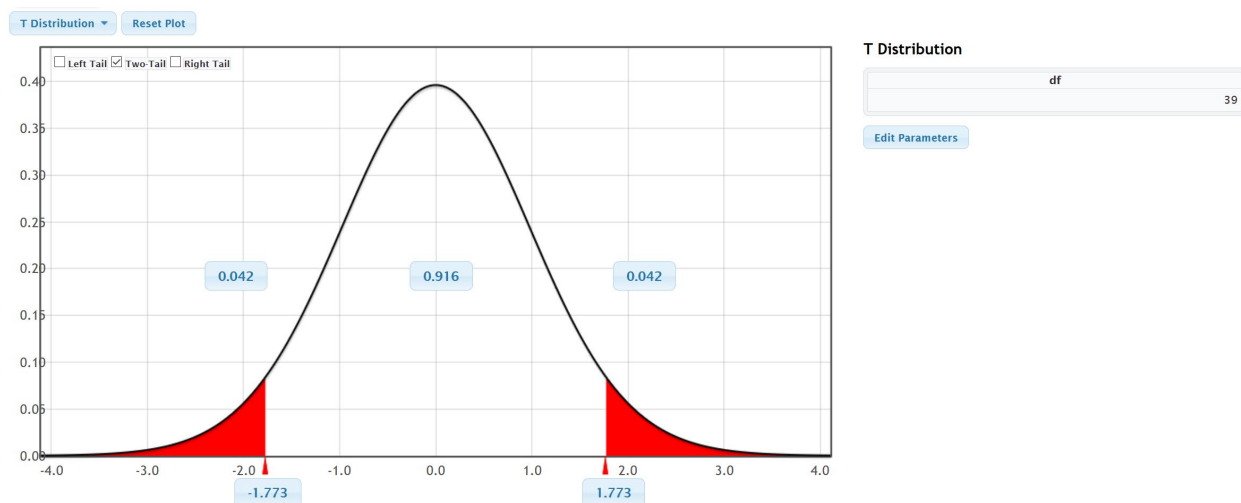


We see that the proportion in the left tail that corresponds to the test statistic cutoff is 0.0009 or 0.09%. This is the estimated P-value, the probability of getting the sample data or more extreme by random chance if the null hypothesis was true.

Example 2: Suppose we are doing a two-tailed hypothesis test that uses the T-test statistic. Remember, our test statistic compares the sample data to the null hypothesis. In this example, our T-test statistic was calculated to be $T=1.773$ and our degrees of freedom was 39. What would the estimated P-value be?

Go to the “theoretical distributions” menu in StatKey at www.lock5stat.com and click on “t”. Under degrees of freedom put in 39 and then click on “Two Tail”. A two-tailed P-value calculation takes a little thought. Notice that there are now two bottom boxes. One in the left tail and one in the right tail. If your T-test statistic is close to the left tail (negative) put in the bottom box in the left tail. If your T-test statistic is close to the right tail (positive) put in the bottom box in the right tail. Since our test statistic is closer to the right tail (positive), we will type in the test statistic of 1.773 into the right bottom box. You do not need to type the test statistic in both boxes. The left tail will automatically adjust to the number you typed in the right tail box. In a two-tailed hypothesis test, “or more extreme” could be any sample data that higher or lower than the parameter in the null hypothesis. So we need to include both of the proportions in the left and right tail. Add the two proportions calculated above the left and right tail. This is the estimated P-value.

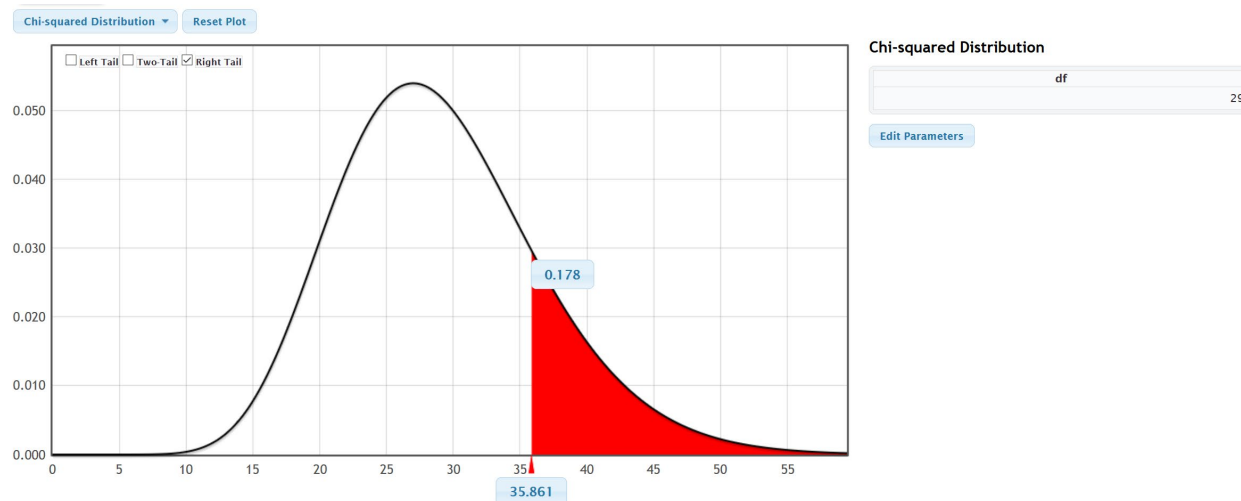




We see that the proportion in the right tail that corresponds to the T-test statistic is 0.042. We also see that the computer has calculated the left tail probability as well. Notice it is also 0.042. So the estimated P-value for this two-tailed hypothesis test is $0.042 + 0.042 = 0.084$ or 8.4%. This is the estimated P-value, the probability of getting the sample data or more extreme by random chance if the null hypothesis was true. Note that a two-tailed P-value is twice as large as a one-tailed P-value from the same data. A common formula that is often used is to take the proportion in the tail corresponding to the test statistic and multiply by two ($0.042 \times 2 = 0.084$).

Example 3: Suppose we are doing a right tailed hypothesis test that uses the chi-squared test statistic. Our test statistic compares the sample data to the null hypothesis. In this example, our chi-squared test statistic was calculated to be $\chi^2 = 35.861$ and our degrees of freedom was 29. What would the estimated P-value be?

Go to the “theoretical distributions” menu in StatKey at www.lock5stat.com and click on “ χ^2 ”. Under degrees of freedom put in 29 and then click on right tail. In the bottom box, put in the test statistic of 35.861. The proportion calculated above is the estimated P-value.



We see that the proportion in the right tail that corresponds to the test statistic cutoff is 0.178 or 17.8%. This is the estimated P-value, the probability of getting the sample data or more extreme by random chance if the null hypothesis was true.



Example 4: Most traditional statistics programs calculate the P-value with this approach. In the previous section on test statistics, we compared the number of alcoholic beverages per week that Math 140 statistics students drink and the number of alcoholic beverages per week that Math 075 pre-statistics students drink. Statcato is using the test statistic, degrees of freedom, and theoretical T-curve to estimate the P-value. Notice that this is a two-tailed hypothesis test with a test statistic of $T=1.846$ and degrees of freedom of 800. We used these numbers with StatKey and got about the same result. In the StatKey printout, we added the tails $0.033 + 0.033 = 0.066$ to get our estimated P-value.

Hypothesis Test - Two population means: confidence level = 0.95

Samples of population 1 in Math 140 alcohol...

Samples of population 2 in Math 075 alcohol...

	N	Mean	Stdev
Population 1	322	2.224	4.684
Population 2	481	1.470	6.884

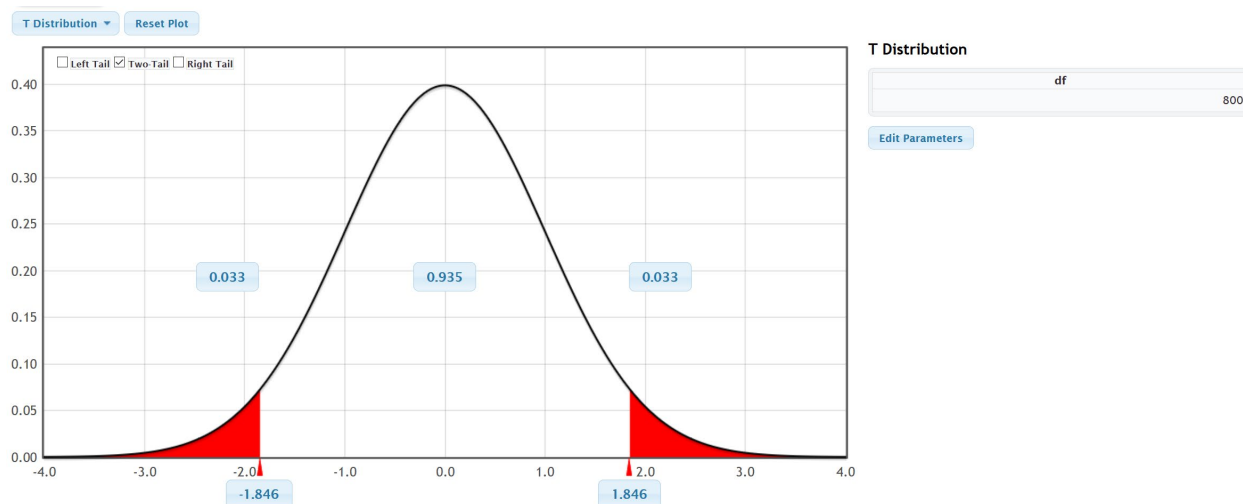
Null hypothesis: $\mu_1 - \mu_2 = 0.0$

Alternative hypothesis: $\mu_1 - \mu_2 \neq 0.0$

* Population standard deviations are unknown. *

DOF = 800

Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.963, 1.963	1.846	0.0653



Method 2: Randomized Simulation (Randomization)

In the previous method, we saw that if we know the test statistic, we can estimate the P-value using a theoretical curve. There are many questions about the accuracy of calculating P-values in this way. For one, we need the data to meet certain assumptions to ensure that the curve is a good approximation of the sampling distribution.

Another approach that is sometimes used to calculate P-value is called “randomized simulation” or “randomization”. This is a more direct way of calculating the P-value. Let’s examine the P-value definition again.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

What would sampling variability look like if the null hypothesis was true? The idea behind randomized simulation is to address this question directly. Computers can create a simulated sampling distribution under the premise that the null hypothesis is true. Once this is accomplished, we can calculate the probability of getting the sample data or more extreme directly. One and two-population hypothesis tests do not require the test statistic calculation and can calculate the P-value directly from either the sample statistic or the difference between the two sample statistics. This technique is also not tied to the accuracy of a theoretical curve, so it has other advantages as well.

Example 1 (Simulation): Open the health data at www.matt-teachout.org and open the men's height data. Some believe that the population mean average height of all men is 68 inches. We want to test the claim that the population mean average height of men is now greater than 68 inches. Here is the null and alternative hypothesis.

$$H_0 : \mu = 68$$

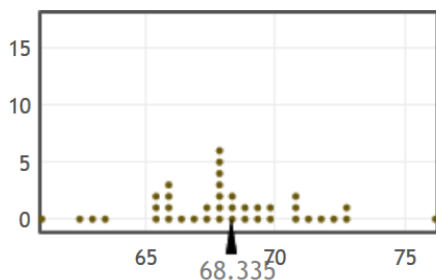
$$H_A : \mu > 68 \text{ (claim)}$$

Notice this is a right tailed test. Go to the "Randomization Hypothesis Tests" menu in StatKey at www.lock5stat.com. Click on "Test for Single Mean". Click on the "Edit Data" button and copy and paste the men's height data into StatKey. Uncheck "Identifier" and check "Header Row". An identifier is a word next to every number that explains something about that value. A header row is a title. Push "Ok". Change the null hypothesis to " $\mu = 68$ ". Now click the "Generate thousand samples" button a bunch of times. We are simulating what sampling variability looks like if the null hypothesis is true. In simulation, it is important to not confuse the simulated samples with the actual original random sample data. The sample mean for the original data is 68.335, so click on "Right Tail" and then put in 68.335 in the bottom box. The proportion above the sample mean is the P-value. Notice we have calculated the probability of getting the sample data or more extreme by sampling variability if the null hypothesis was true. We also did not require the test statistic to calculate it. This is called randomized simulation or randomization.

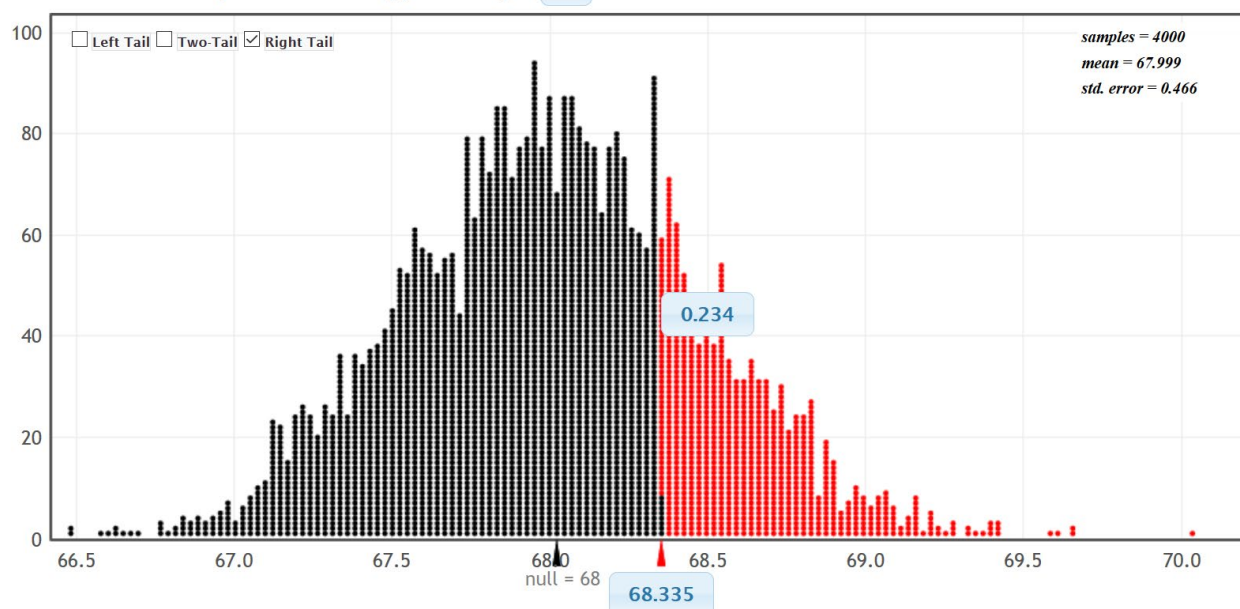
Original Sample

$n = 40$, mean = 68.335

median = 68.3, stdev = 3.02



Randomization Dotplot of \bar{x} . Null hypothesis: $\mu = 68$



We see that the sample mean of 68.335 inches is not in the tail. We also see that the estimated P-value is 0.234 or 23.4%. Both of these tell us that this sample data does not significantly disagree with the null hypothesis and could have occurred because of sampling variability. We will fail to reject the null hypothesis.

Example 2 (Simulation): Suppose we want to compare the percentage (proportion) of math 075 (pre-statistics) students that smoke cigarettes and the percentage of math 140 students that smoke cigarettes. Our claim is that the population proportions are the same. Here is the representative sample data from the Fall 2015 semester at COC and the null and alternative hypothesis.

Math 075 (pre-stat) students: 480 total students, 33 smoke cigarettes

Math 140 (statistics) students: 330 total students, 30 smoke cigarettes

π_1 : Population proportion of pre-stat students that smoke cigarettes at COC.

π_2 : Population proportion of statistics students that smoke cigarettes at COC.

H_0 : $\pi_1 = \pi_2$ (claim)

H_A : $\pi_1 \neq \pi_2$

Notice this is a two-tailed two-population proportion hypothesis test. To use randomized simulation, go to the "Randomization Hypothesis Test" menu and click on "Test for Difference in Proportions". Under the "Edit Data" menu, put in the sample count and total sample size as follows. Since we designated the math 075 pre-stat students as group 1 and math 140 statistics students in group 2, we need to enter the data in that order. Now push "Ok".

Edit data
✕

Please select values for two categories of count and sample size.

Group 1 count:

Group 1 sample size:

Group 2 count:

Group 2 sample size:

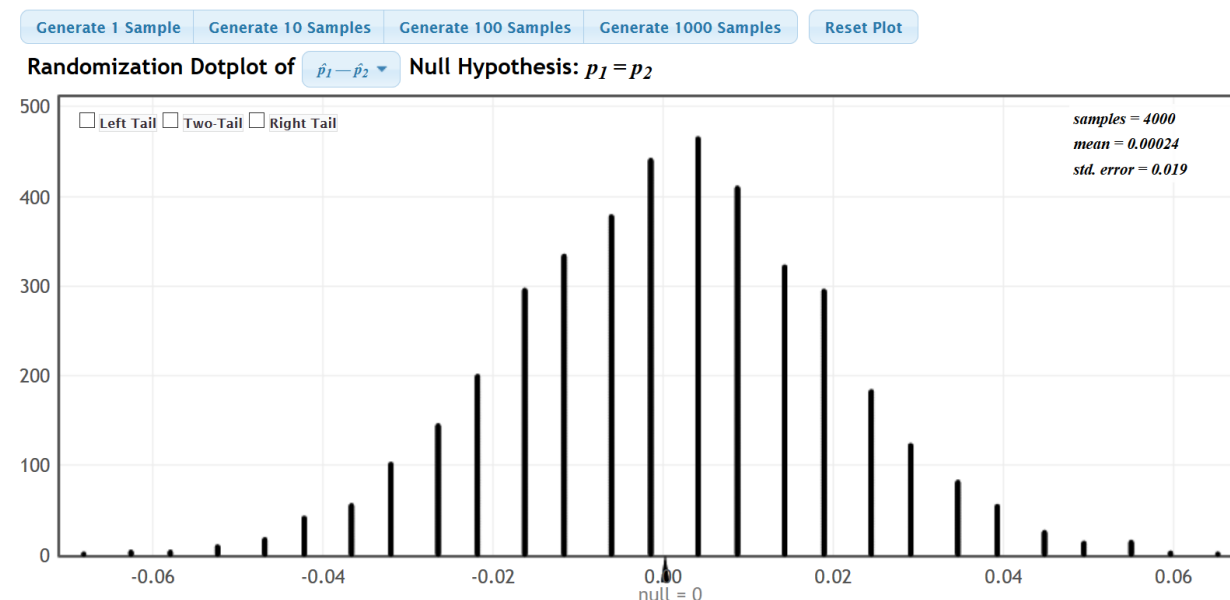


In simulation, it is important not to confuse the original real sample data with all of the simulated samples. Here is the original sample proportions. Notice that the sample proportion for the pre-stat students (\hat{p}_1) was 0.069 and the sample proportion for the stat students (\hat{p}_2) is 0.091. In two-population simulation we will be using the difference between the sample statistics ($\hat{p}_1 - \hat{p}_2$) = -0.022 to calculate the P-value.

Original Sample

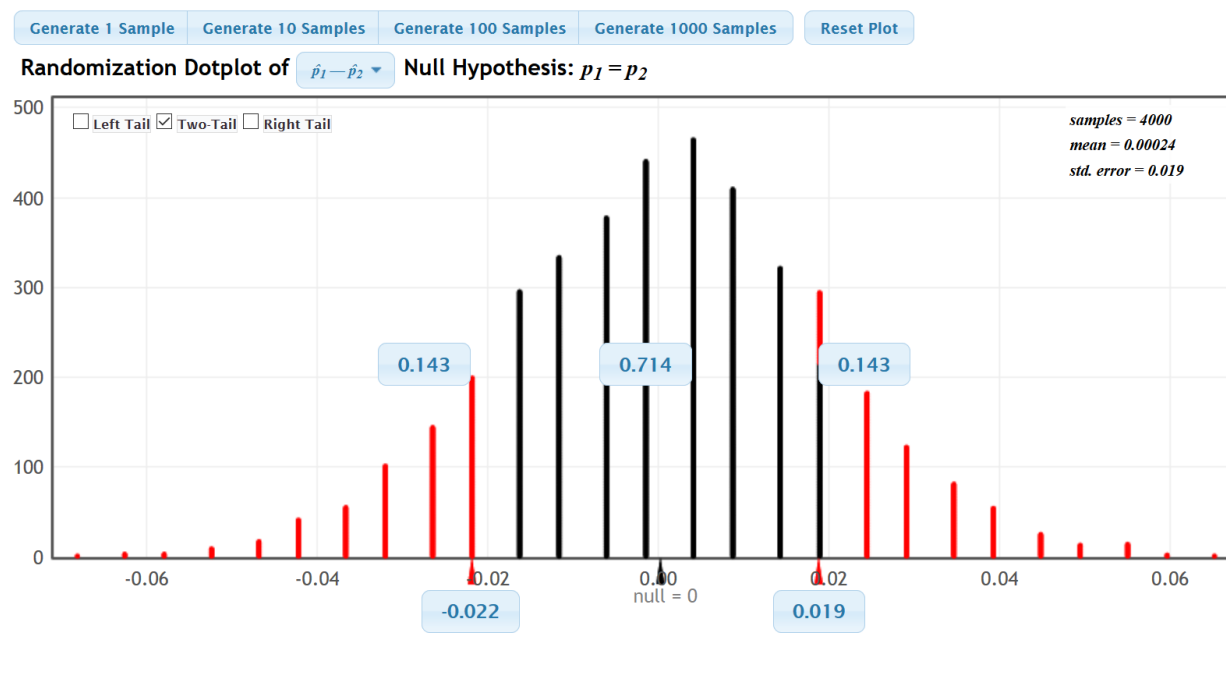
Group	Count	Sample Size	Proportion
Group 1	33	480	0.069
Group 2	30	330	0.091
Group 1-Group 2	3	n/a	-0.022

Let's simulate the null hypothesis. We are creating thousands of random samples from the premise that the populations are equal.



Click on "Two-Tail". Since the difference between the proportions was negative and in the left tail, we will put the difference -0.022 in the left tail. The right tail will automatically adjust. Remember, in a two-tailed hypothesis test, we will need to add the proportions in the top boxes of the two tails to get the estimated P-value. Notice the estimated P-value = 0.143 + 0.143 = 0.286 or 28.6%.





Practice Problems 3C

For #1-32, fill out the following table to interpret the given P-value.

	P-value	P-value %	Significance Level %	Significance Level Proportion	Low P-value or High P-value?	Sample Data significantly disagree with H_0 ? (Yes or No)	Could be sampling variability or Unlikely?	Reject H_0 or fail to reject H_0 ?
1.	0.238		5%					
2.	0.0003		1%					
3.	5.7×10^{-6}		10%					
4.	0.441		5%					
5.	0.138		1%					
6.	0		10%					
7.	0.043		5%					
8.	0.085		1%					
9.	1.4×10^{-4}		10%					
10.	0.112		5%					
11.	0		1%					
12.	0.539		10%					
13.	0.0006		10%					
14.	2.5×10^{-7}		1%					
15.	0.861		5%					
16.	0.199		5%					
17.	0.034		5%					
18.	0.128		1%					
19.	8.6×10^{-4}		10%					
20.	0.0437		5%					
21.	0		1%					
22.	0.612		10%					
23.	0.087		5%					



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

24.	0.0048		1%					
25.	5.5×10^{-7}		10%					
26.	0.0216		5%					
27.	0.444		1%					
28.	0.0539		10%					
29.	0.722		10%					
30.	3.8×10^{-3}		1%					
31.	0.0823		5%					
32.	0.0227		5%					

33. According to a CNN report, 93% of all Americans also own a traditional phone. We disagree with this report. We think that the percentage has decreased as more and more Americans opt to only use a cell phone and throw away their traditional phones. A random sample of 106 Americans was taken and 92 of them owned a traditional phone. The p-value was found to be 0.0168. Use a 5% significance level. The null and alternative hypothesis are given below.

Ho: $\pi = 0.93$

Ha: $\pi < 0.93$ (claim)

- If the null hypothesis is true, what is probability that the sample data or more extreme occurred because of sampling variability?
- Write a standard p-value sentence to describe the true meaning of the p-value in the context of the problem.
- Was the sample data significantly different than the population parameter in H_0 ? Explain why.
- If H_0 was true, could the sample data have occurred because of sampling variability or is it unlikely to be sampling variability? (Could be or unlikely) Explain your answer.
- Use the p-value and the significance level to decide whether we should reject the null hypothesis or fail to reject the null hypothesis. Explain why.

34. According to a recent Newspaper article, the population mean average amount of time people in California spend eating and drinking per day is 1.25 hours. In order to test this claim, we take a random sample of 400 people in California. The average number of hours for the sample was 1.22 and a p-value of 0.248 was found. Use a 10% significance level. The null and alternative hypothesis are given below.

Ho: $\mu = 1.25$ hours (claim)

Ha: $\mu \neq 1.25$ hours

- If the null hypothesis is true, what is probability that the sample data or more extreme occurred because of sampling variability?
- Write a standard p-value sentence to describe the true meaning of the p-value in the context of the problem.
- Was the sample data significantly different than the population parameter in H_0 ? Explain why.
- If H_0 was true, could the sample data have occurred because of sampling variability or is it unlikely to be sampling variability? (Could be or unlikely) Explain your answer.
- Use the p-value and the significance level to decide whether we should reject the null hypothesis or fail to reject the null hypothesis. Explain why.



35. According to an article in *USA Today*, 74% of Americans own a credit card. We disagree with the *USA Today* article. We claim that more than 74% of Americans own a credit card. In order to verify the claim that more than 74% of Americans have a credit card, a random sample of 250 Americans was taken and 77.2% of them owned a credit card and a p-value of 0.1244 was found. Use a 5% significance level. The null and alternative hypothesis are given below.

Ho: $\pi = 0.74$

Ha: $\pi > 0.74$ (claim)

- If the null hypothesis is true, what is probability that the sample data or more extreme occurred because of sampling variability?
- Write a standard p-value sentence to describe the true meaning of the p-value in the context of the problem.
- Was the sample data significantly different than the population parameter in H_0 ? Explain why.
- If H_0 was true, could the sample data have occurred because of sampling variability or is it unlikely to be sampling variability? (Could be or unlikely) Explain your answer.
- Use the p-value and the significance level to decide whether we should reject the null hypothesis or fail to reject the null hypothesis. Explain why.

36. It has long been thought that the population mean average body temperature is 98.6 degrees Fahrenheit. A recent study is now claiming that the population mean average body temperature is really lower than 98.6 degrees. A random sample of 50 adults worldwide was conducted and the average temperature was 98.26 degrees with a p-value of 0.0014 was found. Use a 1% significance level. The null and alternative hypothesis are given below.

Ho: $\mu = 98.6$ degrees Fahrenheit

Ha: $\mu < 98.6$ degrees Fahrenheit (claim)

- If the null hypothesis is true, what is probability that the sample data or more extreme occurred because of sampling variability?
- Write a standard p-value sentence to describe the true meaning of the p-value in the context of the problem.
- Was the sample data significantly different than the population parameter in H_0 ? Explain why.
- If H_0 was true, could the sample data have occurred because of sampling variability or is it unlikely to be sampling variability? (Could be or unlikely) Explain your answer.
- Use the p-value and the significance level to decide whether we should reject the null hypothesis or fail to reject the null hypothesis. Explain why.

37. It has been suggested that at least 10% of the world population is left handed. To test this claim, a sample of 77 randomly selected adults was taken and we found that 11 of them were left handed. A P-value of 0.895 was found. Use a 10% significance level. The null and alternative hypothesis are given below.

Ho: $\pi \geq 0.1$ (claim)

Ha: $\pi < 0.1$

- If the null hypothesis is true, what is probability that the sample data or more extreme occurred because of sampling variability?
- Write a standard p-value sentence to describe the true meaning of the p-value in the context of the problem.
- Was the sample data significantly different than the population parameter in H_0 ? Explain why.



d) If H_0 was true, could the sample data have occurred because of sampling variability or is it unlikely to be sampling variability? (Could be or unlikely) Explain your answer.

e) Use the p-value and the significance level to decide whether we should reject the null hypothesis or fail to reject the null hypothesis. Explain why.

(#38-40) Use the “theoretical distributions” menu in StatKey at www.lock5stat.com to look up the P-value. Click on the button that says “normal”. Click on the tail and enter the test statistic in the bottom box below the tail. Remember in a two-tailed test, you will need to add the proportions in both tails to get the P-value.

38. Z-test statistic = 2.41

Two-tailed test

P-value =

39. Z-test statistic = -1.38

Left-tailed test

P-value =

40. Z-test statistic = 1.02

Right-tailed test

P-value =

(#41-43) Use the “theoretical distributions” menu in StatKey at www.lock5stat.com to look up the following critical values. Click on the button that says “t” and enter the given degrees of freedom. Click on the tail and enter the test statistic in the bottom box below the tail. Remember in a two-tailed test, you will need to add the proportions in both tails to get the P-value.

41. T-test statistic = -2.471

Two-tailed test

Degrees of Freedom = 29

P-value =

42. T-test statistic = 1.352

Right-tailed test

Degrees of Freedom = 34

P-value =

43. T-test statistic = -1.644

Left-tailed test

Degrees of Freedom = 49

P-value =

(#44-45) Use the “theoretical distributions” menu in StatKey at www.lock5stat.com to look up the following critical values. Click on the button that says “ χ^2 ” and enter the given degrees of freedom. Click on the tail and enter the test statistic in the bottom box below the tail. Remember in a two-tailed test, you will need to add the proportions in both tails to get the P-value.

44. χ^2 -test statistic = 38.724

Right-tailed test

Degrees of Freedom = 29

P-value =

45. χ^2 -test statistic = 12.551

left-tailed test

Degrees of Freedom = 39

P-value =



Section 3D – Conclusions

Vocabulary

Population: The collection of all people or objects to be studied.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about “no change”, “no relationship” or “no effect”.

Alternative Hypothesis (H_A or H_1): A statement about the population that does not involve equality. It is often a statement about a “significant difference”, “significant change”, “relationship” or “effect”.

Population Claim: What someone thinks is true about a population.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

Sampling Variability: Also called “random chance”. The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and sampling variability. If the P-value is lower than the significance level, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened because of sampling variability.

Critical Value: We compare a test statistic to this number to determine if the sample data significantly disagrees with the null hypothesis. If the absolute value of the test statistic is higher than the absolute value of the critical value, then the sample data significantly disagrees with the null hypothesis.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value and standard error without a formula.

Conclusion: A final statement in a hypothesis test that addresses the claim and evidence.

Introduction: So far we have learned many of the key ideas used in hypothesis testing. We learned how to write a null and alternative hypothesis and how to determine if the sample data significantly disagrees with the null hypothesis by comparing the test statistic to the critical value. We also learned how to compare the P-value and significance level in order to judge if the sample data occurred by random chance and whether to reject the null hypothesis or fail to reject.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

The last step in a hypothesis test is to write a formal conclusion. A conclusion is a statement about the claim the person made. Do we disagree with that claim (reject the claim) or do we agree with the claim (support the claim)? The conclusion is also a statement about evidence. Do we have evidence to back up our view about the population or not?

Conclusion: A final statement in a hypothesis test that addresses the claim and evidence.

Note: “Reject H_0 ” or “Fail to reject H_0 ” is NOT a conclusion. This is a simple statement of what the P-value tells us about the null hypothesis. It does not address evidence and the claim.

Writing conclusions can be difficult. Many statistics students struggle with the logic of the conclusion. Think of it this way.

Suppose we have a low P-value and H_0 is the claim. A low P-value means we have evidence and will reject H_0 . Since H_0 is our claim, then our conclusion will be that we should “reject the claim”. In other words, we have evidence that the random sample data significantly disagrees with the claim. In this case, our conclusion will be that “we have significant evidence to reject the claim”.

Suppose we have a low P-value and H_A is the claim. Again, a low P-value means that we have evidence and we will reject H_0 . Disagreeing with H_0 is equivalent to agreeing with or supporting H_A . So our conclusion will be that “we have evidence to support the claim”. In other words, random sample data significantly agrees with the claim.

Suppose we have a high P-value and H_0 is the claim. A high P-value means that we do not have evidence and we will fail to reject H_0 . Since H_0 is our claim, then we should think of our conclusion as “failing to reject the claim”. So our conclusion will be that “we do NOT have evidence to reject the claim”. In this case, the random sample data is close to H_0 and does not significantly disagree with the claim.

Suppose we have a high P-value and H_A is the claim. A high P-value is not evidence and means we will fail to reject H_0 . You have to be able to reject H_0 in order to support H_A . Since our random sample data does not significantly disagree with H_0 , it also does not significantly agree with H_A . So our conclusion is that we do not have significant evidence to support the claim. The random sample data does not significantly agree with the claim.

Let's look at some steps for writing a formal conclusion to a hypothesis test.

Step 1: Address the Claim

If the null hypothesis (H_0) is the claim: *There are two possibilities.*

- Yes, we have evidence to reject the claim
- OR
- No, we do not have evidence to reject the claim.

If the alternative hypothesis (H_A) is the claim: *There are two possibilities.*

- Yes, we have evidence to support the claim
- OR
- No, we do not have evidence to support the claim.

Step 2: Address the evidence (Yes or No)

So how do we know if we have evidence or not?



We have already seen that a low P-value less than the significance level indicates significant evidence. Remember, scientists usually require a low P-value as evidence on their reports. If we have a high P-value greater than the significance level, then we do NOT have significant evidence.

Note: While P-value is the usual method for determining significance evidence, a test statistic in the tail also indicates significance evidence. For Z or T-test statistics, if the $|\text{test statistic}|$ is larger than the $|\text{critical value}|$ is considered significant. Remember a large test statistic in the tail will correspond to a small P-value. If the $|\text{test statistic}|$ is smaller than the $|\text{critical value}|$, it is NOT considered significant.

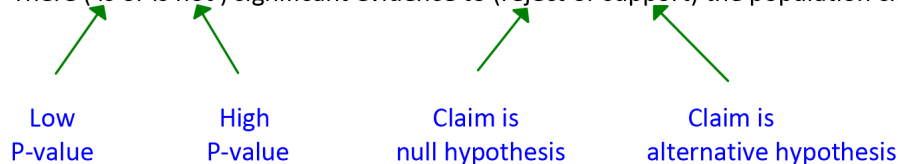
Low P-value (less than significance level) : We have significant evidence.

High P-value (higher than significance level) : We do NOT have significant evidence.

Step 3: Write the conclusion sentence

Remember a low P-value is considered significant statistical evidence but a high P-value is not evidence. When the claim is H_0 , we will either be rejecting or not rejecting the claim. When the claim is H_A , we will either be supporting or not supporting the claim.

There (is or is not) significant evidence to (reject or support) the population claim.



Step 4: Explain the conclusion sentence

Our job as data scientists, statisticians and data analysts is to explain. People rarely understand the language and difficult ideas in statistics. The conclusion is a summary of the hypothesis test, but is rarely understood. It is always good to explain the conclusion in plain language.

There is significant evidence to reject the population claim. *(Random sample data significantly disagrees with the population claim and we have a low P-value as evidence.)*

There is significant evidence to support the population claim. *(Random sample data significantly agrees with the population claim and we have a low P-value as evidence.)*

There is not significant evidence to reject the population claim. *(Unfortunately, the hypothesis test was inconclusive. Random sample data does not significantly disagree with the population claim. We do not have any statistical evidence to disagree with the population claim. The claim might be correct.)*

There is not significant evidence to support the population claim. *(Unfortunately, the hypothesis test was inconclusive. Random sample data does not significantly agree with the population claim. We do not have any statistical evidence to agree with the population claim. The claim might be incorrect.)*

Example 1

A nursing magazine recently claimed that the population mean average amount of a particular medicine that is being given to patients is about 100 milligrams. Looking at a large random sample, we found a P-value of 0.0041 and a 5% significance level ($\alpha = 0.05$) was used in the study. Assuming the data met all the assumptions, what would be the conclusion?

$H_0: \mu = 100$ milligrams (Claim)

$H_A: \mu \neq 100$ milligrams



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a "CC-BY" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Step 1: Address the Claim

The claim is H_0 , so there are two possible conclusions.

- Yes, we have evidence to reject the claim
OR
- No, we do not have evidence to reject the claim.

Step 2: Address the Evidence

The P-value = $0.0041 = 0.41\%$ is less than the 5% significance level. So this is a low P-value close to zero. This means that sampling variability (random chance) is very unlikely and tells us the sample data significantly disagrees with the null hypothesis. We would reject the null hypothesis. Since the claim is the null hypothesis, we have evidence to reject the claim.

Step 3: Writing the conclusion sentence:

"There (*is or is not*) significant evidence to (*reject or support*) the claim."

Notice this problem has a low P-value ("is evidence") and claim is null ("reject").

Formal Statistics Conclusion: "There is significant evidence to reject the claim that the mean average amount of this medicine given to patients is 100 mg."

Step 4: Explain the conclusion in plain language?

Significant statistical evidence disagrees with the claim that the population mean average is 100 mg.

Example 2

An online article is currently estimating that more than 35% of people in the U.S. voted in the last election. Looking at a large random sample, we found a P-value of 0.267 and a 10% significance level ($\alpha = 0.10$) was used in the study. Assuming the data met all the assumptions, what would be the conclusion?

$H_0: p = 0.35$

$H_A: p > 0.35$ (claim)

Step 1: Address the Claim

The claim is H_A , so there are two possible conclusions.

- Yes, we have evidence to support the claim
OR
- No, we do not have evidence to support the claim.

Step 2: Address the Evidence

The P-value = $0.267 = 26.7\%$ is more than the 10% significance level. So this is a high P-value and is not close to zero. This data could have happened because of sampling variability (random chance) and tells us the sample data does not significantly disagree with the null hypothesis. We would fail to reject the null hypothesis. We do not have evidence to reject the null hypothesis. You have to reject the null hypothesis to be able to support the alternative hypothesis. So we do not have evidence to support the claim.

Step 3: Writing the conclusion sentence:

"There (*is or is not*) significant evidence to (*reject or support*) the claim."

Notice this problem has a high P-value ("NOT evidence") and claim is null ("support").



Formal Statistics Conclusion: “There is not significant evidence to support the claim that more than 35% of people in the U.S. voted in the last election.”

Step 4: Explain the conclusion in plain language?

The hypothesis test was inconclusive. Random sample data does not significantly agree with the population claim. We do not have significant statistical evidence to agree with the population claim.

Conclusion Summary (4 Possible Conclusions)

If the claim is H_0 , P-value is low (*Think “Yes Evidence Reject”*)

Conclusion Sentence: There is significant evidence to reject the claim.

(You are rejecting the null hypothesis and the null hypothesis is the claim.)

If the claim is H_0 , P-value is high (*Think “No Evidence Reject”*)

Conclusion Sentence: There is not significant evidence to reject the claim.

(You do not have evidence to reject the null hypothesis and the null hypothesis is the claim.)

If the claim is H_A , P-value is low (*Think “Yes Evidence Support”*)

Conclusion Sentence: There is significant evidence to support the claim.

(*You are rejecting the null hypothesis which means you think the alternative hypothesis is correct and the alternative hypothesis is the claim.*)

If the claim is H_A , P-value is high (*Think “No Evidence Support”*)

Conclusion Sentence: There is not significant evidence to support the claim.

(*You do not have evidence to reject the null hypothesis which means you do not know if the alternative hypothesis is correct and the alternative hypothesis is the claim.*)

Important Notes:

- The claim being the null hypothesis does not in itself mean that you will reject the claim. It may be a “not reject” situation. You would only reject the claim if the claim is H_0 and the P-value was low.
- The claim being the alternative hypothesis does not in itself mean that you will support the claim. It may be a “not support” situation. You would only support the claim if the claim is H_A and the P-value was low.
- Always address the “claim” in a conclusion. Never say that you support H_A or you reject H_0 . That is not a conclusion.



Practice Problems 3D

For #1-30, fill out the following table and write the formal conclusion. A “Low” P-value means that the P-value was lower than the significance level while a “High” P-value means that the P-value was higher than the significance level.

	Claim	P-value	Evidence (Yes or No)	Formal Hypothesis Test Conclusion
1.	H_0	Low		
2.	H_A	High		
3.	H_A	High		
4.	H_0	Low		
5.	H_0	High		
6.	H_A	High		
7.	H_A	Low		
8.	H_0	High		
9.	H_0	Low		
10.	H_A	Low		
11.	H_A	High		
12.	H_0	High		
13.	H_0	Low		
14.	H_A	High		
15.	H_A	Low		
16.	H_0	Low		
17.	H_0	High		
18.	H_A	Low		
19.	H_A	High		
20.	H_0	Low		
21.	H_0	Low		
22.	H_A	High		
23.	H_A	High		
24.	H_0	High		
25.	H_0	Low		
26.	H_A	Low		
27.	H_A	High		
28.	H_0	Low		
29.	H_0	High		
30.	H_A	Low		

31. The hospital claims that less than 4% of people who received the medication showed symptoms of side effects. Use a 1% significance level. (P-value = 0.0027)

Ho: $p \geq 0.04$

Ha: $p < 0.04$ (Claim)

- Should we reject or fail to reject the null hypothesis? Explain why.
- Do we have statistical evidence for our conclusion? Explain why.
- Write the formal hypothesis test conclusion sentence addressing the claim and evidence.
- Explain your conclusion in non-statistics, easy to understand language.



32. We think that the population mean average height of women is at most 63.5 inches.
Use a 5% significance level. (P-value = 0.1843)

Ho: $\mu \leq 63.5$ inches (Claim)

Ha: $\mu > 63.5$ inches

- Should we reject or fail to reject the null hypothesis? Explain why.
- Do we have statistical evidence for our conclusion? Explain why.
- Write the formal hypothesis test conclusion sentence addressing the claim and evidence.
- Explain your conclusion in non-statistics, easy to understand language.

33. Latest polls suggest that the candidate should receive about 54% of the vote.
Use a 10% significance level. (P-value = 0.0711)

Ho: $p = 0.54$ (Claim)

Ha: $p \neq 0.54$

- Should we reject or fail to reject the null hypothesis? Explain why.
- Do we have statistical evidence for our conclusion? Explain why.
- Write the formal hypothesis test conclusion sentence addressing the claim and evidence.
- Explain your conclusion in non-statistics, easy to understand language.

34. The population mean average weight of electrically powered car weighs is more than 2000 pounds. Use a 5% significance level. (P-value = 0.2682)

Ho: $\mu \leq 2000$ pounds

Ha: $\mu > 2000$ pounds (Claim)

- Should we reject or fail to reject the null hypothesis? Explain why.
- Do we have statistical evidence for our conclusion? Explain why.
- Write the formal hypothesis test conclusion sentence addressing the claim and evidence.
- Explain your conclusion in non-statistics, easy to understand language.

35. The medication Toprol is showing real promise in treating migraines. At least 50% of all patients taking Toprol have seen an improvement in their migraine symptoms.
Use a 1% significance level. (P-value = 0.0086)

Ho: $p \geq 0.5$ (Claim)

Ha: $p < 0.5$

- Should we reject or fail to reject the null hypothesis? Explain why.
- Do we have statistical evidence for our conclusion? Explain why.
- Write the formal hypothesis test conclusion sentence addressing the claim and evidence.
- Explain your conclusion in non-statistics, easy to understand language.



36. The population mean average cholesterol for men is different from the population mean average cholesterol for women. Use a 5% significance level. (P-value = 0.0391)

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2 \text{ (Claim)}$$

- Should we reject or fail to reject the null hypothesis? Explain why.
- Do we have statistical evidence for our conclusion? Explain why.
- Write the formal hypothesis test conclusion sentence addressing the claim and evidence.
- Explain your conclusion in non-statistics, easy to understand language.

37. In this state, the population percentage of people that will vote for the democratic candidate will be higher than the percentage of people that will vote for the republican candidate. Use a 10% significance level. (P-value = 0.3144)

$$H_0: p_1 \leq p_2$$

$$H_a: p_1 > p_2 \text{ (Claim)}$$

- Should we reject or fail to reject the null hypothesis? Explain why.
- Do we have statistical evidence for our conclusion? Explain why.
- Write the formal hypothesis test conclusion sentence addressing the claim and evidence.
- Explain your conclusion in non-statistics, easy to understand language.

Section 3E – Type 1 and Type 2 Errors

Vocabulary

Population: The collection of all people or objects to be studied.

Sample: Collecting data from a small subgroup of the population.

Random Sample: Sample data collected in such a way that everyone in the population has an equal chance to be included.

Bias: When sample data does not represent the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about “no change”, “no relationship” or “no effect”.

Alternative Hypothesis (H_A or H_1): A statement about the population that does not involve equality. It is often a statement about a “significant difference”, “significant change”, “relationship” or “effect”.

Population Claim: What someone thinks is true about a population.



Sampling Variability: Also called “random chance”. The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and sampling variability. If the P-value is lower than the significance level, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened because of sampling variability.

Type 1 Error: When biased sample data leads you to support the alternative hypothesis when the alternative hypothesis is actually wrong in the population.

Type 2 Error: When biased sample data leads you fail to reject the null hypothesis when the null hypothesis is actually wrong in the population.

Beta Level (β): The probability of making a type 2 error.

Introduction

It is very difficult to understand large populations. Statisticians and data scientists spend years studying data, sampling variability, sample statistics and population parameters. Yet they can sometimes come to the wrong conclusion about the population. It has nothing to do with their knowledge or ability. The problem is sampling variability. In a sense, data can lead us to a wrong conclusion sometimes. Think about it this way. We are asked to try to understand millions of people in a population, but all we have is one random sample of 250 people. What if that one random sample is biased or is not reflective of what is going on in the population? That random sample can lead us to conclude something is true about the population that really is not true. When this happens, it is not the fault of the statistician or data scientist. They analyzed the data correctly and came to the correct conclusion about the population. The data has lead them astray. These wrong conclusions are called Type 1 or Type 2 errors.

Type 1 Error (False Positive)

A type 1 or type 2 error occurs because the random sample data does not reflect what is really going in the population. A type 1 error is when random sample data gives us a low P-value that is not reflective of the population. Maybe if we collected other random samples they would all have high P-values, but this one biased random sample has lead us to the wrong conclusion.

So biased sample data has given us a low P-value less than the significance level. From this, we rightly make the conclusion that we have evidence to reject the null hypothesis or support the alternative hypothesis. The problem is this the exact opposite conclusion from what is really going in the population. Type 1 errors can look bad on the statistician or data scientist. Remember a type 1 error involves a low P-value and having evidence. Later studies about the population may find that there is not significant evidence. Because the data scientist thinks they have evidence, a type 1 error is sometimes called a “false positive”.

So how can we limit the possibility of making a type 1 error? The probability of making a type 1 error is the significance level (or alpha level “ α ”). So if you want to decrease the chances of making a type 1 error, lower the significance level (lower the alpha level). This is the reason that significance levels are always set low (1%, 5%, or 10%). If the significance level is 5%, then we have only a 5% chance of making a type 1 error. The most common significance level is 5%, but when a statistician is really worried about making a type 1 error, they may lower the significance level to 1% (or a 99% confidence level). Now they have only a 1% probability of making a type 1 error. Making the wrong conclusion about a population can have serious consequences. You may see a statistician decrease the significance level to 0.5% even (99.5% confidence level). In a sense they are making sure the P-value is really close to zero before considering it evidence.



Type 1 Error Summary

- Biased random sample data leads to a low P-value.
- We support the alternative hypothesis when in the population the alternative hypothesis is wrong.
- The probability of a type 1 error is the significance level or alpha level (α)
- To decrease the probability of a type 1 error, decrease the significance level (decrease α).

So why don't statisticians always set the significance level at 1% or 0.5%? Why is the most common significance level 5%? To understand this, we need to take a look at type 2 errors.

Type 2 Error (False Negative)

Type 1 and type 2 errors are inversely related. As the probability of one decreases, the probability of the other tends to increase. So if we decrease the significance level to 1%, the probability of type 1 error will decrease, but the probability of type 2 error will increase. Let's take a look at type 2 errors so we can better understand this.

A type 2 error occurs when random sample data gives a high P-value. The statistician fails to reject the null hypothesis, but later it turns out that this is the wrong conclusion about the population. In the population, the null hypothesis is wrong. This one biased random sample indicated that they did not have evidence, but later studies about the population find that there is evidence. The null hypothesis is actually wrong and the alternative hypothesis is actually correct in the population. Since the data scientist did not have evidence, a type 2 error is sometimes referred to as a "false negative".

So how do we decrease the probability of making a type 2 error. The probability of a type 2 error is called a "Beta Level" (β). Beta levels are highly impacted by sample size. Remember the principle that more data = less error. This is particularly true for type 2 errors. A small sample size usually has a larger probability of making a type 2 error. A small sample size will have a larger beta level. So the main technique for decreasing the probability of making a type 2 error, is to simply collect more data. Increasing the sample size (n) will decrease the probability of making a type 2 error.

Notice that the best way to limit the chances of making a type 1 or type 2 error is to have a large random sample and a small significance level. This is the standard for collecting data.

Type 2 Error Summary

- Biased random sample data leads to a high P-value.
- We fail to reject the null hypothesis when in the population the null hypothesis is wrong.
- The probability of a type 2 error is called the beta level (β).
- To decrease the probability of a type 2 error, increase the sample size. Collect more data.

Setting your Significance and Confidence Levels

Before ever collecting data, a statistician thinks about the consequences of making a type 1 or type 2 error. They set the significance level for the test based on that assessment. Let's look at some of the situations that may lead to using a 1%, 5% or 10% significance level in a hypothesis test.

1% Significance Level ($\alpha = 0.01$): This corresponds to a 99% confidence level. Setting the significance level this low is trying to decrease the probability of type 1 error. At a 1% significance level, the probability of type 2 error will be higher. So this is usually a situation, where someone is trying to avoid a type 1 error, at the expense of allowing a higher probability of type 2 error. A good rule is that if you set the significance level at 1%, collect more random data so that the probability of type 2 error stays relatively low. Collecting more data is not always possible though.

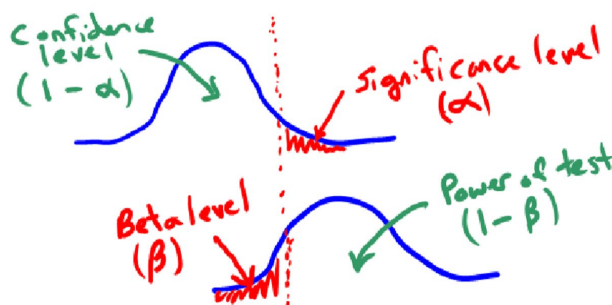
10% Significance Level ($\alpha = 0.10$): This corresponds to a 90% confidence level. Setting the significance level this high will increase the probability of type 1 error. Remember, setting the alpha level high will decrease the beta level. So the probability of type 2 error will be lower. 10% significance levels are sometimes used to decrease the probability of type 2 error when you are not able to collect more data. The scientist is willing to risk the type 1 error, but does not want a type 2 error.



5% Significance level ($\alpha = 0.10$): This corresponds to a 95% confidence level. This is the most common significance and confidence level for a good reason. At a 5% significance level, the alpha and beta levels are both relatively low. So setting the significance level at 5%, keeps the probabilities of type 1 and type 2 errors relatively low.

So why are alpha and beta levels inversely related? Think of the probability distributions associated with a type 1 error is the “alpha curve”. The probability of a type 1 error is the significance level and the complement of the significance level is the confidence level. The probability distribution associated with a type 2 error is the “beta curve”. The complement of the beta level is called the “power” of the hypothesis test. So high confidence will correspond to a low alpha level or a low probability of type 1 error. A hypothesis test with a high power has a low beta level or a low probability of type two error. Remember that larger sample sizes result in a lower probability of type 2 error. As the sample size increases, your hypothesis test power also increases.

Diagram of Alpha levels, Beta levels, confidence levels, and power.



A 1% significance level is like taking the line above and moving it to the right. In the alpha curve on top, the probability in the right tail (significance level) decreases, but look what happens to the beta curve on the bottom. As we pull the line to the right, the beta level increases. Similarly, if we set the significance level at 10% we are pulling the line to the left. The significance level is increasing, but the beta level is now getting smaller.

Example

A pharmaceutical company wants to sell a new medicine in the U.S. To get approval they need to convince the FDA that the medicine is safe and has few side effects. If side effects happen in 3% or more of the people taking the medicine, then the FDA may not approve sale of the medicine in the U.S. If side effects happen in less than 3% of people taking the medicine, then the FDA may approve sale of the medicine in the U.S.

What is the null and alternative hypothesis?

$H_0: p \geq 3\%$ (FDA does not allow medicine to be sold in U.S.)

$H_a: p < 3\%$ (FDA does allow medicine to be sold in U.S.)

Describe the consequences of a type 1 error and what we could do to limit the probability of a type 1 error.

Because of some biased sample data, we got a low P-value and think that the alternative hypothesis is correct when it is not. (In reality, the null hypothesis is correct.) That would mean that the FDA approved sale of the medicine by mistake (false positive). The medicine causes serious side effects in a lot of people. People could die or become very sick. They may sue the pharmaceutical company or the FDA.

This is about as bad of a type 1 error that we can possibly have. To make sure this doesn't happen, we will lower the significance level to 1% (or even lower).

Describe the consequences of a type 2 error and what we could do to limit the probability of a type 2 error.



Because of some biased sample data, we got a high P-value and failed to reject the null hypothesis when in the population, the null hypothesis is really wrong and the alternative hypothesis is correct. So we think that the null hypothesis may be correct when it is not. That would mean that the FDA blocked the sale of a good medicine that rarely causes any side effects. Patients will be deprived of a good medicine and the pharmaceutical company will lose a lot of money in potential profits.

To make sure a type 2 error does not happen, we can increase the sample size. We can collect more data before calculating the P-value and making a decision. We should not increase the significance level just to avoid this type 2 error. We should keep the significance level low because a type 1 error of deaths and side effects is much worse than the pharmaceutical company losing money is.

Practice Problems 3E

Directions: Answer the following questions.

1. Which error is often called a “false negative”?
2. Which error is often called a “false positive”?
3. What is the probability of a type 2 error occurring called?
4. What is the probability of a type 1 error occurring called?
5. What is the complement of the alpha level called?
6. What is the complement of the beta level called?
7. What could we do to decrease the probability that a type 2 error occurs?
8. What could we do to decrease the probability that a type 1 error occurs?
9. If we increase the significance level from 5% to 10% and keep the sample size the same, what will happen to the probabilities for type 1 and type 2 errors?
10. If we decrease the significance level from 5% to 1% and keep the sample size the same, what will happen to the probabilities for type 1 and type 2 errors?
11. What significance level achieves keeps both type 1 and type 2 errors relatively low?
12. Why do type 1 and type 2 errors sometimes occur?
13. If random sample data yielded a high P-value, what kind of error might occur?
14. If random sample data yielded a low P-value, what kind of error might occur?
15. A car company is debating whether to recall its vehicles because of a malfunction in its airbags. Executives think that the defect rate is probably low, but if the airbags malfunction and do not open in 2% or more of crashes, then they will need to put out a general recall. If there are relatively few defective airbags then the company prefers to fix them as needed and not put out a general recall. A random sample of vehicles have their airbags checked for defaults. (The study is currently using a 5% significance level.)

$H_0 : p \geq 0.02$ (Large amount of malfunctioning airbags resulting in recalling all of the vehicles for replacement airbags.)

$H_A : p < 0.02$ (Relatively few malfunctioning airbags resulting in not recalling all of the vehicles for replacement airbags.)

- a) Write a description of a type 1 error and possible consequences of that error in the context of the problem.



- b) Write a description of a type 2 error and possible consequences of that error in the context of the problem.
- c) Would you recommend any changes to the significance level or sample size based on what you know about the type 1 and type 2 errors in this problem? Explain.

16. Mike and his advertisement team have created an advertisement plan for a new flavor of soda. Right now, approximately 16% of soda drinkers are purchasing this flavor. Mike needs to show his bosses that his advertisement plan will increase the percentage of soda drinkers purchasing this new flavor. If Mike's advertising team succeeds in increasing the percentage of customers that prefer this new flavor, then the company will increase supply and make more of the soda to meet demand. If not, then the company will keep the supply as it currently is. After the advertising changes, Mike takes a random sample of customers to determine if the percentage of soda drinkers that like the new flavor has increased. (They are currently using a 5% significance level).

H_0 : $p = 0.16$ (The company will not increase production of the new flavor of soda.)

H_A : $p > 0.16$ (The company needs to increase production of the new flavor of soda to meet the increased demand.)

- a) Write a description of a type 1 error and possible consequences of that error in the context of the problem.
- b) Write a description of a type 2 error and possible consequences of that error in the context of the problem.
- c) Would you recommend any changes to the significance level or sample size based on what you know about the type 1 and type 2 errors in this problem? Explain.

17. Trisha is studying trends in a particular stock to determine if the price of the stock per week will increase or decrease. If the population slope (β_1) in dollars per week is negative, the stock price may decrease and she will recommend selling the stock. If the population slope (β_1) is zero or positive, the stock price may not decrease and she will recommend holding onto the stock.

H_0 : $\beta_1 \geq 0$ (Trisha recommends keeping the stock.)

H_A : $\beta_1 < 0$ (Trisha recommends selling the stock.)

- a) Write a description of a type 1 error and possible consequences of that error in the context of the problem.
- b) Write a description of a type 2 error and possible consequences of that error in the context of the problem.
- c) Would you recommend any changes to the significance level or sample size based on what you know about the type 1 and type 2 errors in this problem? Explain.

18. A global sportswear company is contemplating contributing money to a political candidate in the next election. The managers of the company do not want to contribute unless they are sure the candidate will get the majority of the population vote and win the election. Otherwise, the company will not contribute to the candidates' campaign.

H_0 : $p \leq 0.5$ (The company will not contribute to the political candidates' campaign.)

H_A : $p > 0.5$ (The company will contribute to the political candidates' campaign.)

- a) Write a description of a type 1 error and possible consequences of that error in the context of the problem.
- b) Write a description of a type 2 error and possible consequences of that error in the context of the problem.
- c) Would you recommend any changes to the significance level or sample size based on what you know about the type 1 and type 2 errors in this problem? Explain.



19. A professional basketball team wants to determine whether they should spend millions of dollars to sign a player to a new contract. The team determines that they will sign the player if the population mean average number of points scored per game by the player will be at least twenty. If not, then they will not sign the player.

H_0 : $\mu \geq 20$ (The team will sign the player to a new contract.)

H_A : $\mu < 20$ (The team will not sign the player to a new contract.)

- Write a description of a type 1 error and possible consequences of that error in the context of the problem.
 - Write a description of a type 2 error and possible consequences of that error in the context of the problem.
 - Would you recommend any changes to the significance level or sample size based on what you know about the type 1 and type 2 errors in this problem? Explain.
-

Section 3F – One-population Mean and Proportion Hypothesis Tests

Vocabulary

Population: The collection of all people or objects to be studied.

Sample: Collecting data from a small subgroup of the population.

Random Sample: Sample data collected in such a way that everyone in the population has an equal chance to be included.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about “no change”, “no relationship” or “no effect”.

Alternative Hypothesis (H_A or H_1): A statement about the population that does not involve equality. It is often a statement about a “significant difference”, “significant change”, “relationship” or “effect”.

Population Claim: What someone thinks is true about a population.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

One-Population Proportion Test Statistic (z): The sample proportion is this many standard errors above or below the population proportion in the null hypothesis.

One-Population Mean Test Statistic (t): The sample mean is this many standard errors above or below the population mean in the null hypothesis.

Critical Value: A number we compare our test statistic to in order to determine significance. In a sampling distribution or a theoretical distribution approximating the sampling distribution, the critical value shows us where the tail or tails are. The test statistic must fall in the tail to be significant.



Sampling Variability: Also called “random chance”. The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and sampling variability. If the P-value is lower than the significance level, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened because of sampling variability.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value and standard error without a formula.

Rejecting the null hypothesis: Random sample data significantly disagrees with the null hypothesis.

Failing to reject the null hypothesis: Random sample data does not significantly disagree with the null hypothesis. This does not prove that the null hypothesis is correct however.

Conclusion: A final statement in a hypothesis test that addresses the claim and evidence.

Introduction

So far, we have been learning all of the pieces and ideas behind a hypothesis test. In this section, we will finally start to put the pieces together and perform a hypothesis test from start to finish. While different tests have different test statistics, null and alternative hypotheses, and assumptions, the ideas and methods are very similar. Let's start by looking at the steps for performing a hypothesis test.

Steps to perform a Hypothesis Test

1. Write the null and alternative hypothesis.

Write down the population claim and the null and alternative hypothesis. Consider the type of data you will need and the number of populations to decide the appropriate test for the situation. Is this a right tailed, left tailed or two tailed test? What test statistic should we use? Consider the type 1 and type 2 errors and pick a significance level.

2. Collect random sample data and check the assumptions.

If the sample data has not been collected yet, collect random sample data that is appropriate. Once the sample data is collected, check the assumptions to make sure the data represents the population and can be used for the hypothesis test. You may need to create a histogram to check normality. If certain assumptions are not met, consider revising your hypothesis test method. You may need to move to a non-parametric test or use a randomization technique.

3. Use computer technology to calculate the test statistic, critical value, and P-value and use them to interpret significance and sampling variability.

Compare the test statistic to the critical value. If the test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. Compare the P-value to your chosen significance level. If the P-value is less than the significance level, then it is unlikely to be sampling variability.

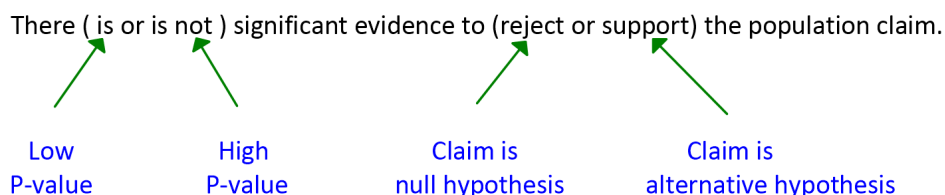
4. Determine if we should reject or fail to reject the null hypothesis.

Remember the P-value is calculated from the assumption that the null hypothesis is true. In a sense, this is about seeing what your P-value tells you about the null hypothesis. If the P-value is less than the significance level, then



we will be rejecting the null hypothesis. If the P-value is higher than the significance level, then we will be failing to reject the null hypothesis. Failing to reject does not mean that the null hypothesis is correct.

5. Write the formal conclusion addressing the claim and evidence.



6. Explain the conclusion in plain language that non-statisticians can understand.

Does the random sample data disagree with claim or agree with the claim? Is it a significant disagreement or a significant agreement? Do we have evidence? What kind of evidence?

Assumptions

In this chapter, we have been discussing test statistics, critical values, theoretical distributions (curves) and P-values. These are the building blocks of hypothesis tests. A key question to think about is how accurate are these building blocks? The standard error is the standard deviation of a sampling distribution. What if that sampling distribution is not normal? The standard error may not be very accurate. To deal with theoretical curves and formulas that involve standard error, the sample data must meet certain assumptions. The good news is that the assumptions for one or two population hypothesis tests are the same as the assumptions we learned for confidence intervals. In our next chapter, we will look at more advanced hypothesis tests where the assumptions get more complicated. Remember the data must meet all of the assumptions for us to use the traditional hypothesis test methods.

Notice that hypothesis tests have some assumptions in common.

Random

One is that the data should be a random sample that is unbiased and represents the population. Statistics students often ask about whether the data can be an unbiased census. The answer is no, but this does not mean that a census is not great data. It is the best. In some ways it is too good. If you have an unbiased census then you know what is going in the population. There is no need for a hypothesis test to surmise what may be true about the population, you know the population parameters. A hypothesis test uses sample data to try to understand population when a census is not possible.

Independence

Another assumption that is particularly difficult to check is that the data values within each sample must be independent of each other. This means that the individuals do not have something in common, at least not for the variable you are studying. This usually rules out the same people measured twice or family members or people from the same classroom or on the same Facebook page. If you know the characteristics of one person, it should not mean that the next person in your data set has a higher probability of that characteristic. A simple random sample from a large population will usually pass the independence criteria. Data scientists usually like the population to be at least ten times larger than the sample size for the individuals to be independent. Think about it. If we collect a simple random sample of 350 people from a population of millions, it is very unlikely that I will accidentally get family members or people that know each other.



Sample Size Requirements

Most assumptions also have some sort of way of checking that the data set is large enough to ensure that it matches the theoretical curve. For proportions, this is usually having at least ten successes and at least ten failures. So ten or more people or objects in the sample data have the categorical criterion you are studying and ten or more do not. If the sample data has at least ten successes and failures, then the sampling distribution for sample proportions is likely to look normal, the standard error will be roughly accurate and the Z-distribution will be a good fit. For mean averages, this is usually having a sample size greater than or equal to 30 or nearly normal. If one of these two criteria is met, then the sampling distribution for sample means is likely to look normal, the standard error will be roughly accurate, and the T-distribution will be a good fit.

Key Question: If we have not collected the data yet, is there a way to know if the sample size is large enough for a given hypothesis test?

Absolutely. If the random sample data has not been collected yet, we can determine what sample sizes should be appropriate. It is generally at least 30 for mean average hypothesis tests, but is more difficult for proportions. We need to collect enough data to ensure that we will get at least ten successes and at least ten failures. Since we have an approximate population proportion (π) in the null hypothesis, we can use this to calculate how many success and failures I am likely to get.

Expected number of successes = $n \times \pi$

Expected number of failures = $n \times (1 - \pi)$

For example, suppose the null hypothesis is $\pi = 0.20$. Will a sample size of 120 randomly selected people be enough?

Expected number of successes = $n \times \pi = 120 \times 0.2 = 24$

Expected number of failures = $n \times (1 - \pi) = 120 \times (1 - 0.2) = 120 \times 0.8 = 96$

Since both of these are greater than ten, a random sample of 120 should be enough.

You can also turn these formulas around and calculate the minimum sample size to get ten. I don't like these too much because you want your expected to be significantly greater than ten. If you calculate for an expected number of success and failures of exactly ten, you may get 8. (We don't always get what we expect in random data.)

In the previous example, here are the minimum sample size calculations.

Minimum sample size (n) based on number of successes and population proportion = $\frac{10}{\pi} = \frac{10}{0.2} = 50$

Minimum sample size (n) based on number of failures and population proportion = $\frac{10}{(1-\pi)} = \frac{10}{0.8} = 12.5$

Always take the larger of the two calculations. In this example, the population proportion was 0.2. We will need a random sample of above 50 if we expect to get at least ten successes and failures.

One-Population Hypothesis Test Assumptions

One-population Mean Assumptions

- The quantitative sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- The sample size should be at least 30 or have a nearly normal shape.

One-population Proportion Assumptions

- The categorical sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least ten successes and at least ten failures.



Example 1

Tim needed to collect some data to test a claim that the population mean average monthly cost of living in California is higher than in Nevada. He put a survey on Facebook. He received data from 217 people living in California and 168 people living in Nevada. The shape of both of the data sets was skewed right. Can we use this data to test perform a two-population mean T-test?

Are the data sets random or representative? *No. This was voluntary response data.*

Are both samples at least 30 or Normal? *Yes. The data sets were not bell shaped (skewed right) but since both of the sample sizes were over 30 (217 and 168) it does pass the 30 or normal requirement.*

Are the data values within the samples independent? *No. These people came from the same Facebook page. They may know each other and even have similar jobs, socioeconomic levels and salaries.*

Are the data values between the samples independent? *No. These California and Nevada samples came from the same Facebook page. They may know each other and even have similar jobs, socioeconomic levels and salaries.*

Since this data did not pass all of the assumptions, we should not use it to judge a population claim with a hypothesis test.

Example 2

Julie needs to test a claim about the percent of people in her hometown that will vote. She took a simple random sample of 45 people and found that 32 said they would vote and 13 said they would not. Can we use this data to test a claim about the population proportion (percentage)?

Is the sample data random or representative? *Yes. This was a simple random sample.*

Are there at least 10 success? *Yes. There are at least 10 success (32 said they will vote)*

Are there at least 10 failures? *Yes. There are at least 10 failures (13 said they will not vote)*

Are the data values within the samples independent? *Yes. There were 45 people out of a relatively large population. So it is unlikely that they would be family members. One person saying they vote will not affect the probability of other people voting. They did come from the same town, but that is the population of interest, so will not be a factor in independence.*

Since the data met the assumptions, we can use this data to test the claim about the population proportion with a hypothesis test and the Z-test statistic. We just need the number of events (voters) = 32 and the total number of trials = 45.

What if the data does not meet all of the traditional assumptions?

Are there ways of working with sample data that does not meet the traditional assumptions?

Absolutely. The one assumption that really needs to be there for any population hypothesis test is that the sample data should be representative, unbiased, and collected randomly. If we are conducting an experiment to prove cause and effect, we will need to separate our groups with random assignment. Besides that, there are many options.

Many advanced techniques in statistics were developed to deal with not meeting all of the traditional assumptions. Non-parametric hypothesis tests were invented for just this purpose. They generally have less assumptions and can be used in a variety of situations. I am a big fan of the non-parametric "Mann-Whitney" test when comparing population averages. Non-parametric hypothesis tests are usually covered in more advanced statistics courses.

One technique that we have discussed is randomized simulation or a randomization test. In some ways, this falls under the umbrella of non-parametric tests. In the section on P-value, we saw that we can create a simulated sampling distribution based on the premise that the null hypothesis is true. We can then calculate the P-value directly



from the simulation. This technique does not require the data to match up with a theoretical distribution curve, so randomization often does not have as many requirements as traditional approaches to hypothesis testing. Notice that there is no requirements for sample size or a normal shape.

Assumptions for a Randomization (Randomized Simulation) Hypothesis Test

- The sample data should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- If multiple samples were collected that were not matched pair, then the data values between the samples should be independent of each other.

Using Technology to Calculate Test Statistics, Critical Values and P-value

Always use technology to calculate test statistics, critical values and P-values. We may need to calculate something from time to time, but always use a computer to do the bulk of the calculations. Your job as a data analyst is to analyze and explain data, not to calculate.

In this book, we have been using two free programs, StatKey and Statcato. Here are some basic directions for using Statcato. Statcato printouts will usually be provided for you to analyze in the problems.

Statcato One-population percentage (proportion) Hypothesis Test with Z-test statistic

Statistics => Hypothesis Tests => 1-Population Proportion

- “Samples in Column” (If you have raw categorical data, enter the column. For example “C1”.)
- “Summarized Sample Data” (Use this option if you have the number of events, total number of trials).
- Put in significance level as a decimal proportion.
- “Hypothesized proportion” (This is asking for the population proportion (π) in H_0)
- “Alternative Hypothesis” (Less than, greater than, not equal. This tells the computer whether to do a left tail, right tail or two-tail hypothesis test.)

Statcato One-population mean average Hypothesis Test with T-test statistic

Statistics => Hypothesis Tests => 1-Population Mean

- “Samples in Column” (If you have raw quantitative data, enter the column. For example “C1”.)
- “Summarized Sample Data” (Use this option if you have the sample size, sample mean and sample standard deviation).
- Put in significance level as a decimal proportion.
- “Hypothesized mean” (This is asking for the population mean (μ) in H_0)
- “Alternative Hypothesis” (Less than, greater than, not equal. This tells the computer whether to do a left-tail, right-tail or two-tail hypothesis test.)
- Leave population standard deviation as “unknown”.

Example 1 (One Population Proportion Z-test) A doctor thinks that the population percent of people in his city that have a certain infection is about 6%. He took a simple random sample of 175 people in the city and found that 13 of them had the infection. Use the following Statcato printout to test the doctor’s claim that exactly 6% have the infection. (Use a 5% significance level.)



Hypothesis Test - One population proportion: confidence level = 0.95

Input: Summary data

Null hypothesis: $p = 0.06$ Alternative hypothesis: $p \neq 0.06$

N	Sample Proportion	Significance Level	Critical Value	Test Statistic Z	p-Value
175	0.074	0.05	-1.96, 1.96	0.796	0.4262

Null and alternative hypothesis?

 $H_0 : \pi = 0.06$ (claim) $H_A : \pi \neq 0.06$

Type of hypothesis test? One population proportion Z-test (two tail)

Assumptions? The data did pass all of the assumptions, so we can proceed with the hypothesis test.

- The sample data was collected randomly.
- The sample data was sufficiently large since it had at least ten success and failures. The data had 13 success and $175 - 13 = 162$ failures.
- Individuals were likely to be independent since it was a simple random sample out of a large population. It is unlikely that the individuals in the sample data will be related.

Test stat $Z = 0.796$

Test Stat Sentence: The sample percent 7.4% was only 0.796 standard errors above the population value 6%.

- Z - Test Statistic does not fall in the tail and so is not significant. (The sample proportion needs to be higher than +1.96 or more to be in the right tail (significantly higher) or -1.96 or less to be in the left tail (significant lower.)
- This tells us that the sample value (7.4%) was not significantly different from the population value (6%)
- The sample data does not significantly disagree with the null hypothesis.

P-value = 0.426

P-Value Sentence: If H_0 is true, and the population percent really is 6%, we had a 42.6% probability of getting the sample percentage of 7.4% or more extreme by random chance.

- This is a high P-value. (The P-value of 42.6% is much larger than the 5% significance level.)
- If H_0 is true, this tells us that the sample data could have happened by random chance (sampling variability).
- Sample value of 7.4% is not significantly different from the population value of 6%.
- There is not a significant disagreement between the sample data and the null hypothesis.

Reject H_0 or Fail to reject H_0 ? Fail to reject H_0 since the P-value is larger than the significance level.

Conclusion?

There is not significant evidence to reject the claim that 6% of the population have the infection.

(The random sample data does not significantly disagree with the claim that 6% of the city is infected. The doctor might be correct, but we do not have evidence.)



Practice Problems Section 3F

1. Write out the traditional assumptions for a one-population proportion Z-test.
2. Write out the traditional assumptions for a one-population mean average T-test.
3. Write out the assumptions for a one-population randomized simulation hypothesis test.

(#4-7) Directions: Check the assumptions for a one-population proportion Z-test. You must check all of the assumptions. Explain your answers. Does the problem meets all the assumptions? Can we use this data to perform the hypothesis test?

4. Marsha works for the Republican Party and is asked to test a claim about the percentage of people in Sacramento will vote for the republican candidate in the next election. She has a computer randomly pick phone numbers with a Sacramento area code. She then calls the phone numbers and asks people if they would vote for the republican candidate. She spoke with 123 people and 37 said they would vote for the republican candidate.

5. A health organization is doing a study on smoking tobacco and need to test a claim about the population percentage of tobacco smokers use a pipe. They had people hang out in stores that sell tobacco and pipes and counted how many total people came to the store and how many of them used a pipe. They found that out of the 79 people, eight of them used a pipe.

6. The COC Admissions department needs to test a claim about the population percentage of students that would be in favor of using a new program to register for classes. They put a link on their website so that any students that want to try out the program can. The students can then take a survey and say how well they like the new system. A total of 247 students tried the system and 112 of them said they liked the new system.

7. Michelle, a teacher at Valencia High, wants to test a claim about the population percentage of students at Valencia High school will be attending COC. She gives the students in her English 1 class a questionnaire to fill out that asks where they will be attending college. Of the 34 students in her class, 26 are planning to attend COC.

(#8-11) Directions: Check the assumptions for traditional one-population mean average T-test. You must check all of the assumptions. Explain your answers. Does the problem meets all the assumptions? Can we use this data to perform the hypothesis tests?

8. A company wants to test a claim about the population mean average amount of alcohol drunk by people vacationing in Las Vegas per day. They also want to test a claim about the population standard deviation. They posted a survey on Facebook asking how much people drink when on vacation in Las Vegas. A total of 8,355 people responded and listed how much they drink. The sample mean was 3.5 drinks per day with a standard deviation of 1.2 drinks. A histogram of the alcohol consumption data was skewed to the right.

9. Jimmy works for a company that designs new homes and just moved to Oklahoma City. Jimmy's boss wants him to test a claim about the population mean average price of all homes in Oklahoma City and test a claim about the population variance. Jimmy had a computer randomly select 28 homes. He then found out the price paid for each home. A histogram of the home prices showed a normal distribution. The sample mean average price of the 28 homes was \$212.4 thousand dollars with a standard deviation of \$27.6 thousand dollars.

10. Rick works for a sports equipment manufacturing company. He needs to test a claim about the population mean average amount customers spend at his stores and test a claim about the population standard deviation. He went to the store by his house and kept track of how much people spent on a Tuesday. There were 45 customers and the histogram of the data showed a skewed right distribution. The mean average price of the sample was \$71 with a standard deviation of \$19.

11. Mike is trying to take an opinion poll and found out the average amount of money in thousands of dollars that people in Los Angeles would pay per year in order to have an NFL football team. He randomly selects three streets in Los Angeles and asks every person living on those streets. His sample size was 63 people, but the histogram was drastically skewed right. The mean average amount of money was 1.3 (thousand dollars) with a standard deviation of 0.6 (thousand dollars).



(#12-24) Directions: For each of the following problems answer the following questions.

a) Give the null and alternative hypothesis. Which is the claim? Is this a right-tailed, left-tailed, or two-tailed test? Explain how you know what tail to use.

b) Check the assumptions.

c) Give the test statistic and write the standard sentence to explain it. Compare your test statistic to the critical value. Did the sample data significantly disagree with the null hypothesis? Explain how.

d) Give the p-value and write the definition sentence to explain it. Could the sample data have happened because of sampling variability (random chance) or is it unlikely to be sampling variability? Explain why.

e) Compare the p-value to the significance level. State whether you reject the null hypothesis or fail to reject the null hypothesis. Explain your answer.

f) Write the standard conclusion.

g) Explain your conclusion in easy to understand language.

12. The United States has the highest teen pregnancy rate in the industrialized world. The Center for Disease control says that as of 2011, 33% of girls get pregnant before the age of 20. We are wondering if the teen pregnancy rate this year is even higher than 33%. A random sample of 400 girls is taken. Of the 400 girls randomly selected, 144 of them were pregnant before the age of 20. Use the following Statcato printout and a 5% significance level to test the claim that the teen pregnancy rate is higher than 33%.

N	Sample Proportion	Significance Level	Critical Value	Test Statistic Z	p-Value
400	0.36	0.05	1.645	1.276	0.1010

13. According to the National Association of College Stores, digital textbooks are projected to account for approximately 13% of course materials sold by the fall of 2012. In a random sample of 260 college course materials, 39 (or 15%) of the sample course materials were digital. Use the random sample data, a 10% significance level, and StatKey to test the claim that 13% of course materials are digital. Under the "Randomization Hypothesis Tests" menu, click on "Test for Single Proportion". Under the "Edit Data" button put in the random sample data of 39 (count) and 260 (sample size). Change the null hypothesis to 13% (0.13). Determine if the sample proportion fell in the tail and use the sample proportion (0.15) to calculate the P-value. After performing the simulation, make note of the standard error and use the following formula to calculate the test statistic. Write a sentence to explain the test statistic.

$$Z\text{-test statistic} = \frac{\text{Sample Proportion} - \text{Population Proportion}}{\text{Standard Error}}$$

14. An online source suggests that one out of every three people in the U.S have high blood pressure and the population proportion of U.S. adults is 33.3%. Another website disagrees with this and claims that the true percentage of U.S. adults with high blood pressure is actually dramatically lower than 1 in 3 (33.3%). A random sample of 500 U.S. adults found that 165 of them had high blood pressure. Use the Statcato printout below and a 10% significance level to test the claim that less than 33% of U.S. adults have high blood pressure.

N	Sample Proportion	Significance Level	Critical Value	Test Statistic Z	p-Value
500	0.33	0.10	-1.282	-0.142	0.4434



15. Childhood obesity has more than tripled in the past 30 years. The percentage of children aged 6–11 years in the United States who were obese increased from 7% in 1980 to nearly 20% in 2008. If this trend continues, we can expect that the percent of young children that are obese today to be significantly greater than 20%. In order to test this claim, a random sample of 800 children in the U.S. was taken and 179 of them were found to be obese. Use StatKey and a 10% significance level to test the claim that the population proportion of obese children in the U.S. is higher than 0.20. Under the “Randomization Hypothesis Tests” menu, click on “Test for Single Proportion”. Under the “Edit Data” button put in the random sample data of 179 (count) and 800 (sample size). Change the null hypothesis to 20% (0.20). Determine if the sample proportion fell in the tail and use the sample proportion (0.22375) to calculate the P-value. After performing the simulation, make note of the standard error and use the following formula to calculate the Z-test statistic. Write a sentence to explain the test statistic.

$$\text{Z-test statistic} = \frac{\text{Sample Proportion} - \text{Population Proportion}}{\text{Standard Error}}$$

16. Test the claim that exactly 25% of Math 140 statistics students take their class at the COC Canyon Country Campus. A census of the fall 2015 semester indicated that of the 334 statistics students at COC, 111 of them took their class at the Canyon Country campus. Assume the population of interest is all COC statistics students from all semesters. Use the Statcato printout below and a 5% significance level to test the claim.

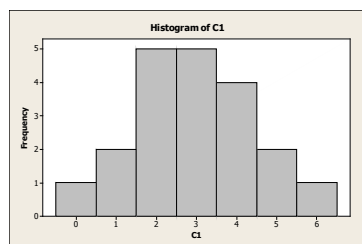
N	Sample Proportion	Significance Level	Critical Value	Test Statistic Z	p-Value
334	0.332	0.05	-1.96, 1.96	3.475	0.0005

17. Use StatKey and a 5% significance level to test the claim that more than 50% of Math 075 pre-stat students identify as female. A census of the fall 2015 semester indicated that of the 481 pre-stat students at COC, 270 of them identified as female. Assume the population of interest is all COC statistics students from all semesters. Under the “Randomization Hypothesis Tests” menu, click on “Test for Single Proportion”. Under the “Edit Data” button put in the random sample data of 270 (count) and 481 (sample size). Change the null hypothesis to 50% (0.50). Determine if the sample proportion fell in the tail and use the sample proportion (0.5613) to calculate the P-value. After performing the simulation, make note of the standard error and use the following formula to calculate the Z-test statistic. Write a sentence to explain the test statistic.

$$\text{Z-test statistic} = \frac{\text{Sample Proportion} - \text{Population Proportion}}{\text{Standard Error}}$$

18. The manager at a local Starbucks wants to make sure that customers wait less than 4 minutes from the time they order to the time that they pick up their coffee. In order to test this, twenty random customers were selected and the staff measured the number of minute between when the person ordered and when their drink was ready. The sample mean was 2.870 minutes and the sample standard deviation was 1.379 minutes. Here is a histogram of the twenty wait times. Does this data meet the assumptions necessary to perform a hypothesis test? If so, use a 1% significance level to test the claim that the average wait time is less than 4 minutes.

N	Sample Mean	Stdev s	Significance Level	Critical Value	Test Statistic	p-Value
20	2.87	1.379	0.01	-2.540	-3.665	0.0008



19. Redwood trees are the tallest plants on Earth. California is famous for its giant Redwood trees. However, just how tall are they? A random sample of 47 California Redwood trees was taken and their heights measured. (This was not easy by the way.) The sample mean average height was 248 feet with a standard deviation of 26 feet. Does this data meet the assumptions necessary to perform a hypothesis test? If so, use a 5% significance level to test the claim that Redwood trees have an average height greater than 240 feet.

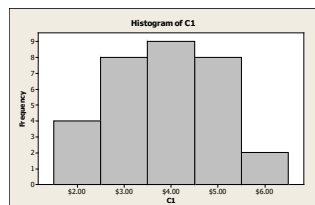
N	Sample Mean	Stdev s	Significance Level	Critical Value	Test Statistic	p-Value
47	248.0	26.0	0.05	1.679	2.109	0.0202

20. Maria is planning to attend UCLA. She is curious what the average age of UCLA students is. Since most students that attend UCLA are in their 20's yet there are also students up to 70 years old, the sample data is skewed right. The college conducted a random sample of 65 students and found that the sample mean was 29.0 years old with a standard deviation of 5.2 years. Does this data meet the assumptions necessary to perform a hypothesis test? If so, use a 10% significance level to test the claim that the average age of students at UCLA is 30 years old.

N	Sample Mean	Stdev s	Significance Level	Critical Value	Test Statistic	p-Value
65	29.0	5.2	0.10	-1.669, 1.669	-1.550	0.1260

21. Mike wants to know the average price of a hamburger. He randomly selected 24 restaurants and recorded the price of a regular hamburger. The sample mean price was \$3.88 and the sample standard deviation was \$1.14. A histogram of the data is below. Does this data set meet the assumptions necessary to perform a hypothesis test? If so, use a 10% significance level to test the claim that the average price of a hamburger is greater than \$3.50?

N	Sample Mean	Stdev s	Significance Level	Critical Value	Test Statistic	p-Value
24	3.88	1.14	0.10	1.320	1.633	0.0580



22. Use StatKey at www.lock5stat.com and the "Math 140 Survey Data Fall 2015" at www.matt-teachout.org to test the claim that the population mean average age of math140 students is higher than 21 years old. Check the assumptions and use a 1% significance level. Under the "Randomization Hypothesis Test" menu, click on "Test for Single Mean". Under "Edit Data" copy and paste the age data into StatKey and click on "raw data". Do not click the "identifier" button. Push "OK". What is the sample mean age of the statistics students? Create the randomized simulation and make a note of the standard error. Use the significance level to determine the tail. Does the sample mean fall in the tail? Use the sample mean and the simulation to calculate the P-value. Finish the hypothesis test. After your test is complete, use the following formula to calculate the T-test statistic. Write a sentence to explain the T-test statistic.

$$\text{T-test statistic} = \frac{\text{Sample Mean} - \text{Population Mean}}{\text{Standard Error}}$$



23. Use StatKey at www.lock5stat.com and the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org to test the claim that the population mean average weight of math140 students is less than 160 pounds. Check the assumptions and use a 5% significance level. Under the “Randomization Hypothesis Test” menu, click on “Test for Single Mean”. Under “Edit Data” copy and paste the weight data into StatKey and click on “raw data”. Do not click the “identifier” button. Push “OK”. What is the sample mean weight of the statistics students? Create the randomized simulation and make a note of the standard error. Use the significance level to determine the tail. Does the sample mean fall in the tail? Use the sample mean and the simulation to calculate the P-value. Finish the hypothesis test. After your test is complete, use the following formula to calculate the T-test statistic. Write a sentence to explain the T-test statistic.

$$\text{T-test statistic} = \frac{\text{Sample Mean} - \text{Population Mean}}{\text{Standard Error}}$$

24. Test the claim that the population mean average height of math140 students is 64 inches. Assume the data met the assumptions. Use a 10% significance level.

Use StatKey at www.lock5stat.com and the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org to test the claim that the population mean average height of math140 students is 64 inches. Check the assumptions and use a 10% significance level. Under the “Randomization Hypothesis Test” menu, click on “Test for Single Mean”. Under “Edit Data” copy and paste the height data into StatKey and click on “raw data”. Do not click the “identifier” button. Push “OK”. What is the sample mean height of the statistics students? Create the randomized simulation and make a note of the standard error. Use the significance level to determine the tail. Does the sample mean fall in the tail? Use the sample mean and the simulation to calculate the P-value. Finish the hypothesis test. After your test is complete, use the following formula to calculate the T-test statistic. Write a sentence to explain the T-test statistic.

$$\text{T-test statistic} = \frac{\text{Sample Mean} - \text{Population Mean}}{\text{Standard Error}}$$

Chapter 3 Review Sheet

Topics:

- Use population claims to construct the null and alternative hypothesis. Also, know how to determine the type of tail for the test.
- Know the assumptions (conditions) necessary to do a one-population mean or proportion hypothesis test.
- Be able to explain the meaning of the Z-test statistic for one-population proportion hypothesis tests.
- Be able to explain the meaning of the T-test statistic for one-population mean average hypothesis tests.
- Be able to determine how likely it is for the sample data to have occurred by sampling variability (random chance).
- Know how to interpret significance by judging whether the test statistic falls in the tail corresponding to the critical value.
- Know how to interpret significance by judging whether the sample statistic falls in the tail corresponding to the significance level in a randomized simulation.
- Know how to use “Theoretical Distributions” menu in StatKey to use the significance level to look up critical values.
- Know how to use “Theoretical Distributions” menu in StatKey to use the test statistic to look up the P-value.
- Know how to create randomized simulations and use the significance level to judge the tails.
- Know how to create randomized simulations with StatKey and use the sample statistic to calculate the P-value.
- Know how to use the P-value and significance level to determine if the sample data could have happened because of sampling variability or if it was unlikely.
- Know the definition of P-value.



- Know how to use the P-value and significance level to determine if we should reject the null hypothesis or fail to reject the null hypothesis.
- Be able to write the formal conclusion for a hypothesis test and explain its meaning.
- Know the definitions and terms associated with Type 1 and Type 2 errors. Know how to decrease the chances of having a Type 1 or Type 2 error.

Chapter 3 Vocabulary Terms and Definitions

Population: The collection of all people or objects to be studied.

Sample: Collecting data from a small subgroup of the population.

Random Sample: Sample data collected in such a way that everyone in the population has an equal chance to be included.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about “no change”, “no relationship” or “no effect”.

Alternative Hypothesis (H_A or H_1): A statement about the population that does not involve equality. It is often a statement about a “significant difference”, “significant change”, “relationship” or “effect”.

Population Claim: What someone thinks is true about a population.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

One-Population Proportion Test Statistic (z): The sample proportion is this many standard errors above or below the population proportion in the null hypothesis.

One-Population Mean Test Statistic (t): The sample mean is this many standard errors above or below the population mean in the null hypothesis.

Critical Value: A number we compare our test statistic to in order to determine significance. In a sampling distribution or a theoretical distribution approximating the sampling distribution, the critical value shows us where the tail or tails are. The test statistic must fall in the tail to be significant.

Sampling Variability: Also called “random chance”. The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and sampling variability. If the P-value is lower than the significance level, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened because of sampling variability.



Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value and standard error without a formula.

Type 1 Error: When biased sample data leads you to support the alternative hypothesis when the alternative hypothesis is actually wrong in the population.

Type 2 Error: When biased sample data leads you fail to reject the null hypothesis when the null hypothesis is actually wrong in the population.

Beta Level (β): The probability of making a type 2 error.

Conclusion: A final statement in a hypothesis test that addresses the claim and evidence.

Chapter 3 Review Notes

Steps to Writing the Null and Alternative Hypothesis

1. **Write down the population claim in symbolic notation.**
 (Note: This can be the null or the alternative hypothesis.)
 (Note: This is the statement the scientist thinks is true.)
 (Note: In one-population tests, always write the parameter on the left and the number on the right.)
2. **Write down the opposite of the population claim.**
 (Note: This can be the null or the alternative hypothesis.)
 (Note: This is the statement that would be true if the scientist is wrong.)
 (Note: In one-population tests, always write the parameter on the left and the number on the right.)
3. **The statement with equality ($=, \leq, \geq$) is always the null hypothesis (H_0).**
 (Note: The null hypothesis is usually " $=$ ". Statements with " \leq " or " \geq " are often changed to " $=$ ".)
4. **The statement that does NOT have equality ($\neq, <, >$) is always the alternative hypothesis (H_A).**

Which Tail should we use in a one-population Hypothesis Test?

- **Right-Tailed Test:** Alternative Hypothesis (H_A) is **GREATER THAN ($>$)**
- **Left-Tailed Test:** Alternative Hypothesis (H_A) is **LESS THAN ($<$)**
- **Two-Tailed Test:** Alternative Hypothesis (H_A) is **NOT EQUAL (\neq)**



One-Population Hypothesis Test Assumptions (Conditions)

Assumptions for One-Population Mean Hypothesis Test

- Random Sample Data
- Individuals Independent
- Sample data normal or sample size at least 30

Assumptions for One-Population Proportion (%) Hypothesis Test

- Random Sample Data
- Individuals Independent
- Sample data has at least 10 successes and at least 10 failures.

Assumptions for One-Population Randomized Simulation Hypothesis Test

- Random Sample Data
- Individuals Independent

Summary Table of P-Value, Test Statistic, Simulation, Significance, Sampling Variability and Evidence

	SIGNIFICANT	NOT SIGNIFICANT
	Test Statistic falls in tail determined by a critical value	Test Statistic does NOT fall in tail determined by a critical value
	OR	OR
	Low P-value ($P\text{-value} \leq \text{significance level}$)	High P-value ($P\text{-value} > \text{significance level}$)
	OR	OR
	Sample Statistic falls in tail determined by the significance level of randomized simulation.	Sample Statistic does NOT fall in tail determined by the significance level of randomized simulation.
Is the sample data significantly different than H_0 ?	Significantly Different	NOT Significantly Different
Could the sample data happen because of sampling variability (random chance) if H_0 is true?	Unlikely	Could Happen



Reject H_0 or Fail to Reject H_0 ?		Reject H_0	Fail to Reject H_0
Is there significant evidence?		Significant Evidence	NOT Significant Evidence

Summary Table: Conclusions

	Claim is H_0	Claim is H_A
Low P-value ($P\text{-value} \leq \text{significance level}$)	There is significant evidence to reject the claim. (Evidence indicates that the claim may be wrong.)	There is significant evidence to support the claim. (Evidence indicates that the claim may be correct.)
High P-value ($P\text{-value} > \text{significance level}$)	There is NOT significant evidence to reject the claim. (Random sample data does not disagree with the claim. The claim could be true, but we do not have evidence.)	There is NOT significant evidence to support the claim. (Random sample data disagrees with the claim. The claim could be wrong, but we do not have evidence.)

Chapter 3 Review Practice Problems

- Write a definition for the following key terms.
 - hypothesis test
 - Null hypothesis
 - Alternative Hypothesis
 - Population Claim
 - Test statistic
 - one-population proportion Z test statistic
 - one-population mean T test statistic
 - Critical Value
 - Sampling Variability (Random Chance)
 - P-value
 - significance level (alpha level)
 - Randomized Simulation
 - beta level
 - type 1 error
 - type 2 error
 - Conclusion
- How is randomized simulation used in hypothesis testing and describe what it can tell us.
- How can we know if the sample data significantly disagrees with the null hypothesis?



4. How can we determine how likely it is for the sample data to have occurred because of sampling variability?
5. How do we know if we reject the null hypothesis or fail to reject the null hypothesis?
6. What are the four steps to writing conclusions?
7. What assumptions do we need to check for the following:

When testing a hypothesis about a one-population mean average?

When testing a hypothesis about a one-population proportion (percentage)?

When testing a hypothesis about one-population using randomized simulation?

8. Fill out the following table regarding test statistics and critical values.

Test Statistic	Critical Value	Does sample significantly disagree with H_0 or not?
$T = +1.774$	± 2.751	
$Z = -2.481$	-1.96	
$T = -3.394$	± 2.566	
$Z = +1.362$	$+1.645$	

9. Fill out the following table regarding P-value and Significance levels.

P-value	P-value %	Significance Level	Sampling Variability or Unlikely	Reject H_0 or Fail to reject H_0 ?
0.0002		5%		
0.3327		1%		
1.84×10^{-5}		10%		
0.0941		5%		

10. Fill out the following table to practice writing conclusions.

P-value	Claim	Write the Conclusion addressing Evidence and claim
Low	H_0	
High	H_A	
High	H_0	
Low	H_A	

11. Write the null and alternative hypotheses for the following. Which is the claim? Is this a left-tailed, right-tailed, or two-tailed test?

- a) "We used to think that the population mean average normal body temperature for all humans was exactly 98.6 degrees Fahrenheit. Now we think it is lower."
- b) "The percentage of all people in China with Typhoid (π_1) used to be higher than the percentage of all people in India with Typhoid (π_2). Now we think the percentage is about the same."
- c) "The population mean average age of students at UCLA (μ_1) is the same as the population mean average age of students at USC (μ_2)."

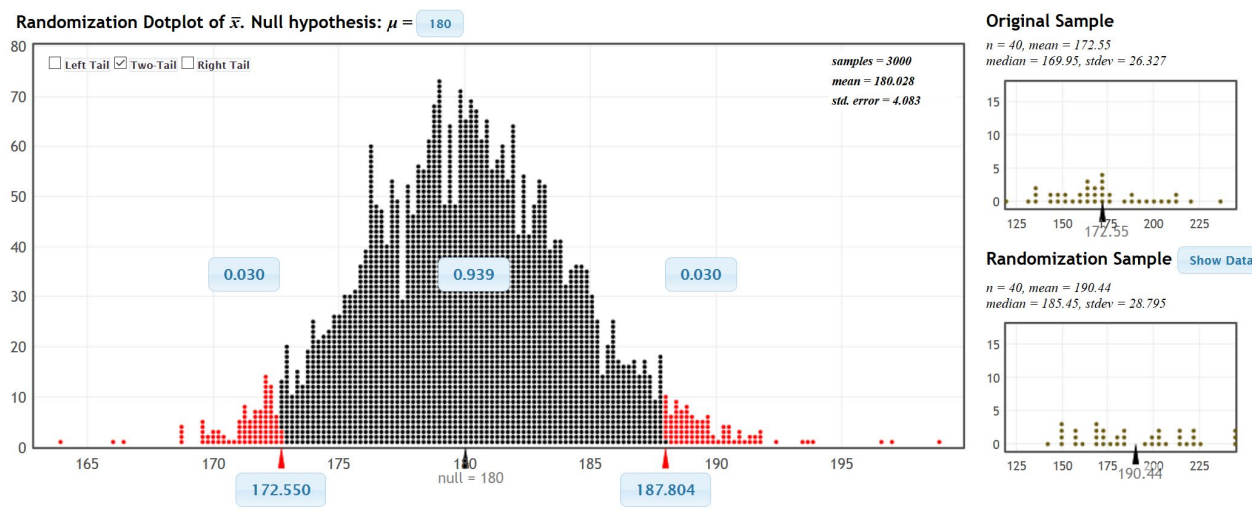
12. Answer the following questions about type 1 and type 2 errors.

- a) What is a type 1 error?
- b) What is a type 2 error?
- c) What is the probability of type 1 error called?
- d) What is the probability of type 2 error called?
- e) Why do type 1 and type 2 errors occur?
- f) What can we do to limit the chances of a type 1 error?
- g) What can we do to limit the chances of a type 2 error?



- h) Which significance level is best for keeping both type 1 and type 2 errors low?
 i) If the significance level is 1%, what happens to the probability of type 1 and type 2 errors?
 j) If the significance level is 10%, what happens to the probability of type 1 and type 2 errors?

13. An article states that the mean average weight of men in the U.S. is 180 pounds. The random health data at teachoutcoc.org was pasted into StatKey and the following randomized simulation was created in order to test the article's claim. Use a 1% significance level. Make sure to check the assumptions. Give the null and alternative hypothesis, estimate the P-value, and the conclusion. Write a sentence to explain the P-value. Was there a significant difference between the sample mean and the population mean? How likely was it that the sample data occurred by random chance if the population mean really is 180 pounds?

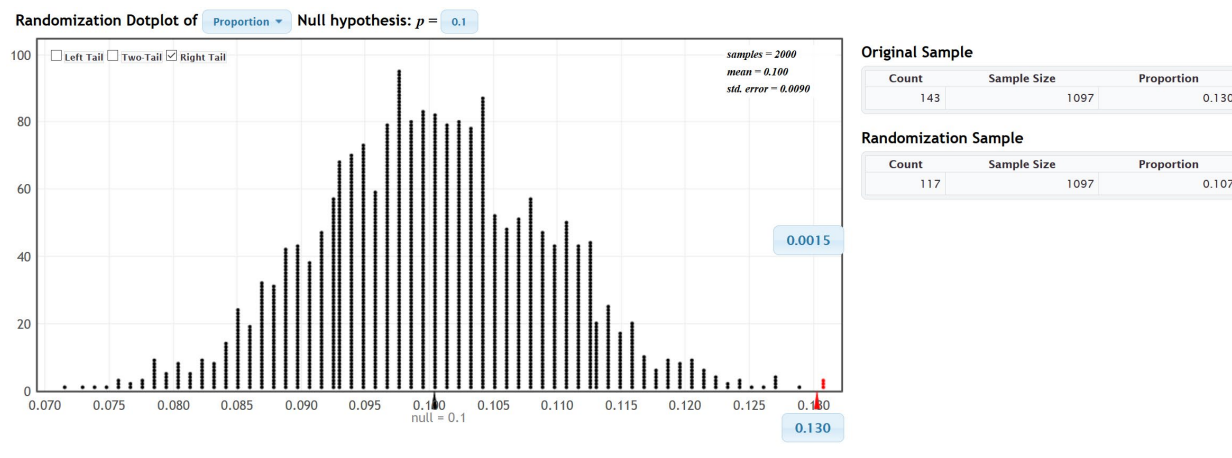


14. Use the Statcato printout below to test the following claim. A speaker at a nursing convention said that the population mean average hourly salary of a registered nurse used to be \$25 per hour, but now is greater than \$25 per hour. A random sample of 28 nurses gave a sample mean of \$26.82 and a standard deviation of \$4.37. A histogram of the salaries showed a bell shaped distribution. Use a 5% significance level. Make sure to check the assumptions. Give the null and alternative hypothesis, test statistic, P-value, and the conclusion. Write a sentence to explain the test statistic. Write a sentence to explain the P-value. Was there a significant difference between the sample mean and the population mean? How likely was it that the sample data occurred by random chance if the population mean really is \$25?

N	Sample Mean	Stdev s	Significance Level	Critical Value	Test Statistic	p-Value
28	26.82	4.37	0.05	1.703	2.204	0.0181



15. Use simulation on www.lock5stat.com to test the following claim. The Harris Poll conducted a random survey in which they asked 1097 women "How many tattoos do you currently have?" Of the 1,097 females surveyed, 143 responded that they had at least one tattoo. A tattoo magazine claimed that more than 10% of women have at least one tattoo. Use a 5% significance level to test the magazine's claim. Give the null and alternative hypothesis, the estimated P-value, and the conclusion. Write a sentence to explain the P-value. Was there a significant difference between the sample percent and the population value? How likely was it that the sample data occurred by random chance from a 10% population.



Chapter 4: Categorical and Quantitative Relationship Tests

Vocabulary

Categorical Data: Another word for qualitative data. Data that is generally in the form of labels that tell us something about the people or objects in the data set. For example, the country a person lives in, the person's occupation, type of pet, or smoking status.

Quantitative Data: Numerical measurement data. The data is made up of numbers that measure or count something and have units. Also taking an average of the data should make sense.

Random Sample: Collecting data from a population in such a way that every person in the population has an approximately equal chance of being chosen. This technique tends to give us data with less sampling bias.

Random Assignment: Take a group of people or objects and randomly put them into two or more groups. This is a technique used in experiments to create similar groups. Similar groups help to control confounding variables so that the scientist can prove cause and effect.

Hypothesis Test: A procedure for testing a claim about a population.

Random Chance: Another word for sampling variability. The principle that random samples from the same population will usually be different and give very different statistics.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data and the number of populations.

Critical Value: If the test statistic is higher than this number, then the sample data significantly disagrees with the null hypothesis. The z or t score critical values are also used to calculate margin of error in confidence intervals.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

P-value: The probability of getting the sample data or more extreme by random chance if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. If the P-value is lower than this number, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened by random chance. This is also the probability of making a type 1 error.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value without a formula.

Introduction: In the last chapter, we introduced the idea of a hypothesis test. This is a procedure for checking a claim about a population. People make claims all the time about populations. In the last chapter, we introduced the one-population hypothesis test to check a claim about a specific population. This last chapter will continue the discussion of hypothesis testing. A very common hypothesis test is determine if population variables may be related or not. The type of variable is very important though. We cannot analyze a categorical/categorical relationship the same way we analyze a quantitative/quantitative relationship.

There is a common thread in all of these relationship hypothesis tests that is very important to understand from the outset. If we find that a population parameter is the same in various groups (populations), then it does not seem to matter what group we are in, we get about the same thing. This would indicate that the variable that decides the groups is not related to the parameter we are studying. Alternatively, if the population parameter is significantly different in various groups (populations), then it does matter what group we are in. This would indicate that the variable that decides the groups is related to the parameter we are studying. Therefore, the null hypothesis will usually be “not related” or “independent” because we will need to show parameters are equal in various populations. The alternative hypothesis will be “related” or “associated” because this corresponds to parameters being different or not equal. Remember equal is always the null hypothesis.

H_0 : The variables are NOT related (not associated, independent) – *parameters from various populations are equal*

H_A : The variables are related (associated, dependent) – *parameters from various populations are not equal*

Note about cause and effect: Remember just because you prove two variables are related does not imply that one causes the other. In chapter 1, we learned that to prove cause and effect we need to control confounding variables with experimental design. When a scientist needs to prove cause and effect, they will often use random assignment instead of a random sample to control the confounding variables.

Section 4A – Categorical/Quantitative Relationships: Two Population Mean Hypothesis Test

Suppose we want to determine if categorical variables are related to a quantitative variable or not. A common technique would be to examine the population means from the various groups determined by the categorical variable. If the population means from the quantitative data are equal in the groups, then that would indicate that the categorical variable that determines the groups is not related to the quantitative variable. It did not matter what group we are in, since the means are about the same. If the mean averages for the groups are significantly different, then it does matter what group we are in. This would indicate that they are related. For this section, we will be focusing on categorical data with only two options. This leads to a two-population mean average hypothesis test. If the categorical data has three or more variables, then that would lead to an ANOVA test. We will cover that test in our next section.



Important note: Just because variables are related does not imply cause and effect. To prove cause and effect, we need to use experimental design.

Null and Alternative Hypotheses

Here are common null and alternative hypotheses for the two-population mean average hypothesis test. Notice equal (not related) is the null hypothesis and not equal (related) is the alternative hypothesis.

Let us suppose that the groups are independent of each other. There are a couple different ways of writing the null and alternative hypothesis. Notice that saying that the population means are equal is the same as saying the difference is zero. A not equal alternative hypothesis would be a two-tailed test.

μ_1 : Mean Average of Population 1

μ_2 : Mean Average of Population 2

(Two-tailed, two-population mean from independent groups)

$H_0 : \mu_1 = \mu_2$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_1 \neq \mu_2$ (categorical variables are related to the quantitative variable)

OR

$H_0 : \mu_1 - \mu_2 = 0$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_1 - \mu_2 \neq 0$ (categorical variables are related to the quantitative variable)

We can also specify that the population mean of population 1 is higher or lower than population 2. Notice that still indicates that the categorical and quantitative variables are related. If the alternative hypothesis is less than, then it is a left tailed test. Less than points to the left. If the alternative hypothesis is greater than, then it is a right tailed test. Greater than points to the right. While some people prefer to use " \leq " or " \geq " symbol for the null hypothesis, I generally do not. Mainly because of the relationship idea. The null hypothesis is not related which must be equal to.

(Right-tailed, two-population mean from independent groups)

$H_0 : \mu_1 = \mu_2$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_1 > \mu_2$ (categorical variables are related to the quantitative variable)

$H_0 : \mu_1 - \mu_2 = 0$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_1 - \mu_2 > 0$ (categorical variables are related to the quantitative variable)

(Left-tailed, two-population mean from independent groups)

$H_0 : \mu_1 = \mu_2$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_1 < \mu_2$ (categorical variables are related to the quantitative variable)

$H_0 : \mu_1 - \mu_2 = 0$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_1 - \mu_2 < 0$ (categorical variables are related to the quantitative variable)

Sometimes we may have the same people measured twice or some one-to-one pairing between the groups. When this happens, we call this "matched pairs". If you recall from our discussion of matched pair confidence intervals, we subtract the ordered pairs. This creates the difference column of data. We then calculate the mean and standard deviation of the differences.



μ_d : Mean Average of Differences between the populations

(Two-tailed, two-population mean from matched pair data)

$H_0 : \mu_d = 0$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_d \neq 0$ (categorical variables are related to the quantitative variable)

(Right-tailed, two-population mean from matched pair data)

$H_0 : \mu_d = 0$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_d > 0$ (categorical variables are related to the quantitative variable)

(Left-tailed, two-population mean from matched pair data)

$H_0 : \mu_d = 0$ (categorical variables are not related to the quantitative variable)

$H_A : \mu_d < 0$ (categorical variables are related to the quantitative variable)

Two-population Mean Hypothesis Test Assumptions

The assumptions for two-population hypothesis tests are the same as for two-population confidence intervals that we discussed in previous chapters. The assumptions are slightly different depending on if the groups are matched pair or independent. Two-population hypothesis tests are also used in experimental design. In that case, we need the groups to be randomly assigned in order to control confounding variables. Another way to control confounding variables is to measure the same group of people twice (matched pair).

Two-population Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape.

Two-population Mean Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

Two-population mean T-test statistic

The two population mean T-test statistic is very similar to the two-population proportion Z-test statistic. It just compares two sample means instead of two sample proportions. The two-population mean T-test statistic is used to determine if two sample means are significantly different. It can also be thought of as determining if the difference between the sample means is significantly different from zero or some other difference in the null hypothesis.

It is important not to confuse one and two-population test statistics. Recall that the one-population mean T-test statistic counts the number of standard errors that the sample mean (\bar{x}) is above or below the population mean (μ) in the null hypothesis. If the T-test statistic is positive, then the sample mean (\bar{x}) is a certain number of standard errors



“above” the population mean (μ). If the T-test statistic is negative, then the sample mean (\bar{x}) is a certain number of standard errors “below” the population mean (μ).

The two-population mean T-test statistic will count how many standard errors that the sample mean for group 1 (\bar{x}_1) is above or below the sample mean for group 2 (\bar{x}_2). If the T-test statistic is positive, it is “above”. If the T-test statistic is negative, it is “below”. The two-population T-test statistic can also be thought of as the number of standard errors that the difference between the means is from zero or some other claimed difference.

Here are a couple of different formulas used by computer programs. If you recall in our discussions of confidence intervals two-population mean comparisons can come from data that is independent groups (like men and women) or matched pairs (like the same people measured twice). Again, it is not important for you to calculate these by hand with a calculator. Computers do the heavy lifting. Focus on being able to explain the test statistic and using it to determine significance.

These formulas are much easier to calculate if you already know the standard error. For independent groups, “ n_1 ” is the sample size of group 1 and “ n_2 ” is the sample size of group 2. The standard deviation for group 1 is “ s_1 ” and the standard deviation for group 2 is “ s_2 ”. For matched pairs, the sample sizes of both groups are the same (n). The mean of the differences between the matched pairs is “ \bar{d} ” and the standard deviation of the differences is “ s_d ”.

(Independent Groups) Two-population mean T-test statistic

$$T = \frac{(\text{Sample Mean for group 1 } (\bar{x}_1) - \text{Sample Mean for group 2 } (\bar{x}_2))}{\text{Standard Error}} \quad \text{OR} \quad \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left[\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)\right]}}$$

(Matched Pairs) Two-population mean T-test statistic

$$T = \frac{(\text{Mean of Differences between matched pairs } (\bar{d}))}{\text{Standard Error}} \quad \text{OR} \quad \frac{(\bar{d})}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example

Suppose we want to test the claim that the level of statistics student at COC is not related to the amount of alcohol they drink. If they are not related, then the population mean average amount of alcoholic beverages per week between COC pre-stat students should be the same as the mean average amount of alcoholic beverages per week between COC statistics students. In this case, the claim is the null hypothesis. Notice these are independent groups. Population 1 is COC statistics students and population 2 is COC pre-stat students. Here is the null and alternative hypothesis. Notice this will be a two-tailed hypothesis test. Use a 5% significance level.

$H_0 : \mu_1 = \mu_2$ (The level of stat student is not related to the amount of alcohol beverages per week) (Claim)

$H_A : \mu_1 \neq \mu_2$ (The level of stat student is related to the amount of alcohol beverages per week)

OR

$H_0 : \mu_1 - \mu_2 = 0$ (The level of stat student is not related to the amount of alcohol beverages per week) (Claim)

$H_A : \mu_1 - \mu_2 \neq 0$ (The level of stat student is related to the amount of alcohol beverages per week)

The data can be found on www.matt-teachout.org. We will be using the “COC Statistics Survey Data Fall 2015”. We will be comparing the number of alcoholic drinks for Math 140 (statistics) students to the number of alcoholic drinks for Math 075 (pre-stat) students.

Let us start by checking the assumptions. The two data sets are not matched pair so we will check the assumptions for independent groups.

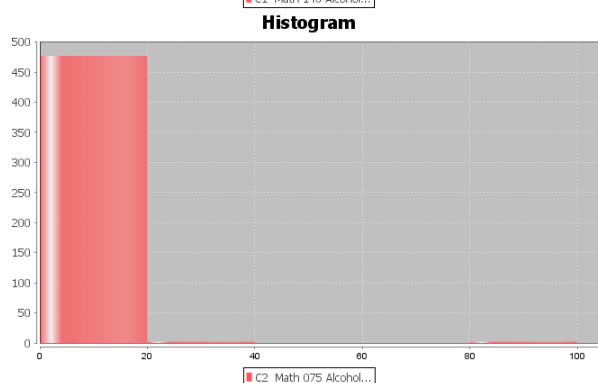
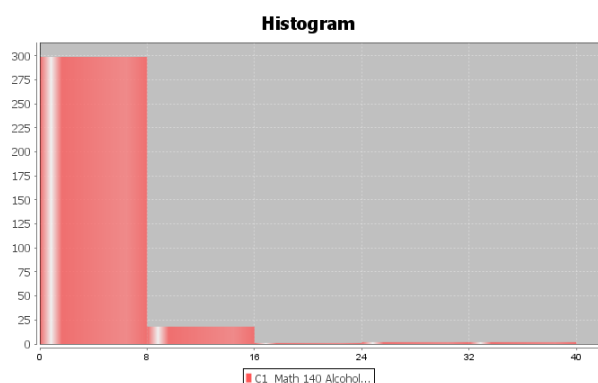


Assumptions for Two-Population Mean (Independent Groups)

- Sample data should be collected randomly or represent the population. If it is an experiment, then the groups should be randomly assigned. **Yes. This sample data was not random, but it was a census of the fall 2015 semester, so is likely to be representative of COC students.**
- The sample sizes for both groups should be at least 30 or nearly normal. **Yes. The sample sizes are 322 and 481, which are both over 30. Even though both data sets are skewed right, it still passes the at least 30 or normal requirement.**

Descriptive Statistics

Variable	N total
C1 Math 140 Alcoholic Beverages per Week	322
C2 Math 075 Alcoholic Beverages per Week	481



- Data values within the samples and between the samples should be independent of each other. **No. Some of the Math 140 students and Math 075 students may have come from the same class. Groups of friends may have similar alcohol consumption.**

We want to use Statcato to calculate the test statistic, critical value and degrees of freedom. Since these data sets are over 300, we will need to add few rows before copy and pasting the data into Statcato. These data sets have a sample size of 322 and 481, so we will add about 200 rows to Statcato. Go to the “Edit” menu in Statcato and click on “Add multiple rows/columns”. Put in 200 next to “rows” and push OK. We will get our sample data sets from the “Math 075 Survey Data Fall 2015” and the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org under the “statistics” menu and “data sets”. Copy and paste the alcoholic beverages per week data for both groups into two columns of Statcato. Since these are independent groups, we will go to the “Statistics” menu in Statcato, click on “Hypothesis Tests”, and then click on “2-population means”. Since our raw data is in two columns, click on “Samples



in two columns". Type in the column for math 140 alcoholic beverages under population 1 and math 075 as population 2. Notice saying that the groups are equal is the same as saying the difference is zero. Therefore, the "hypothesized mean difference" should be zero. In addition, the alternative hypothesis is "Not Equal To" and significance level is 0.05. Push OK.

Hypothesis Test: 2-Population Means

Help F1

Inputs

☐ Samples in one column
Labels in column:
Values in column:

☒ Samples in two columns
Population 1:
Population 2:

☐ Summarized sample data

	Sample Size	Mean	Standard Deviation
Population 1:	<input type="text"/>	<input type="text"/>	<input type="text"/>
Population 2:	<input type="text"/>	<input type="text"/>	<input type="text"/>

Population Standard Deviations/Variances

☐ Population standard deviations known
 σ_1 :
 σ_2 :

☐ Assume population variances are equal

Alternative Hypothesis

Alternative Hypothesis:
Hypothesized Mean Difference:

Significance

☒ Significance Level: 0 - 1.00 (e.g. 0.05)
☐ Confidence Level: 0 - 1.00 (e.g. 0.95)

OK Cancel

Here is the Statcato printout.

Hypothesis Test - Two population means: confidence level = 0.95

Samples of population 1 in Math 140 alcohol...

Samples of population 2 in Math 075 alcohol...

	N	Mean	Stdev
Population 1	322	2.224	4.684
Population 2	481	1.470	6.884

Null hypothesis: $\mu_1 - \mu_2 = 0.0$

Alternative hypothesis: $\mu_1 - \mu_2 \neq 0.0$

* Population standard deviations are unknown. *

DOF = 800

Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.963, 1.963	1.846	0.0653

Let us write a sentence to explain the T-test statistic. Remember, in this case group one is Math 140 statistics students and group 2 is Math 075 pre-statistics students. The sample mean number of alcoholic beverages per week for group 1 (\bar{x}_1) is 2.224 and the sample mean number of alcoholic beverages per week for group 2 (\bar{x}_2) is 1.47. Also, note that the test statistic is positive, indicating the group 1 is above group 2.

Sentence to explain the T-test statistic: The sample mean average number of alcoholic beverages per week for Math 140 statistics students is 1.846 standard errors above the sample mean average number of alcoholic beverages per week for Math 075 pre-statistics students.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Is it significant?

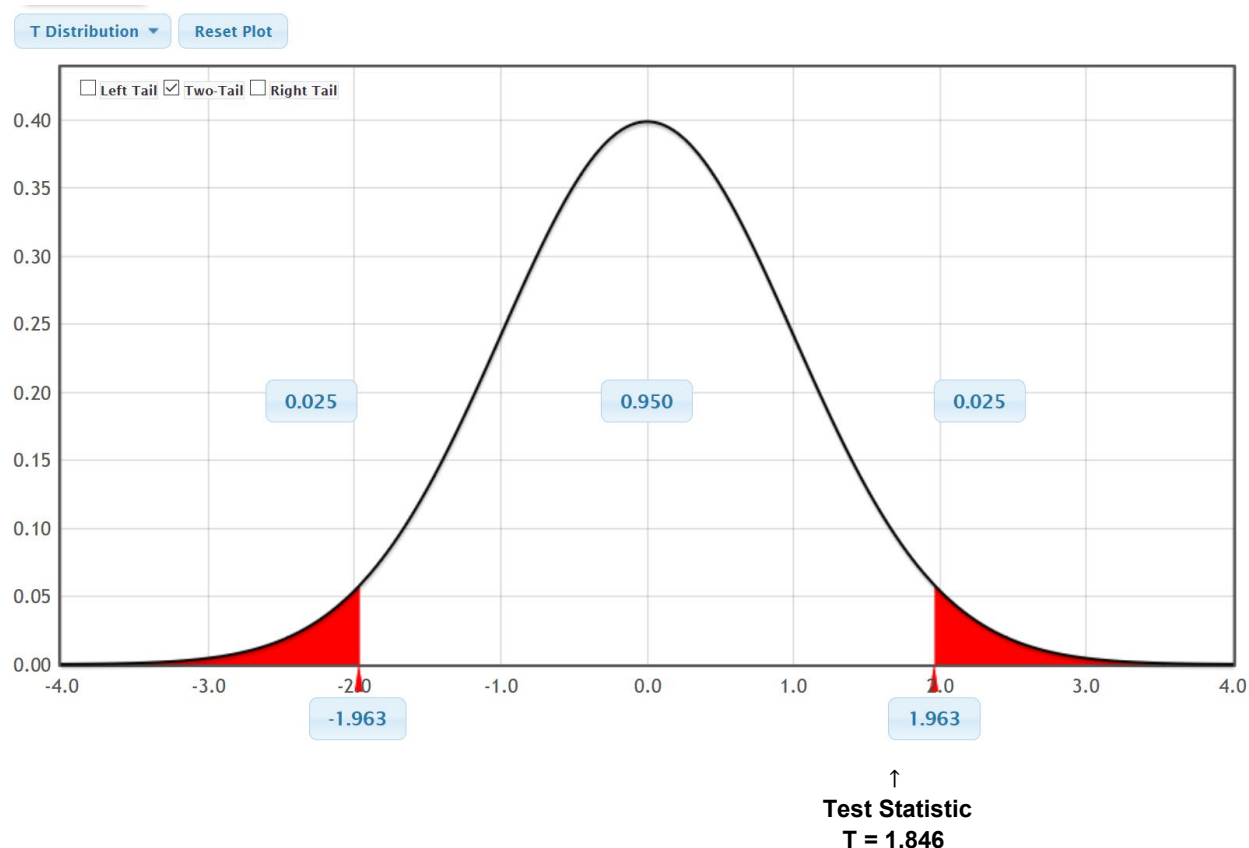
Notice that the test statistic did not fall in one of the tails determined by the critical values. This indicates that the sample means are not significantly different. This also indicates that the sample data does not significantly disagree with the null hypothesis.

Notice that the degrees of freedom is 800 in the Statcato printout. How did Statcato calculate this? The formula for two-population mean degrees of freedom from independent groups is given below. You will need the sample sizes and standard deviations for both groups. Again, never calculate this by hand. Statcato calculated it for us. If you do not have Statcato, there are many two-population mean degrees of freedom calculators for independent groups. Here is one I like to use. (<http://web.utk.edu/~cwieck/TwoSampleDoF>). You will want to round the degrees of freedom to the ones place. In this example, the app above gave “800.7819” which is close to what Statcato gave. We usually round the degrees of freedom down in order to account for more variability. An easier formula for two-population mean degrees of freedom for independent groups is to use the smaller of $n_1 - 1$ or $n_2 - 1$.

$$(\text{Independent groups}) \text{ Two-population mean degrees of freedom} = \frac{\left[\left(\frac{s_1^2}{n_1} \right) + \left(\frac{s_2^2}{n_2} \right) \right]^2}{\left[\left(\frac{s_1^2/n_1}{n_1 - 1} \right) + \left(\frac{s_2^2/n_2}{n_2 - 1} \right) \right]}$$

(Matched Pair) Two-population mean degrees of freedom = $n - 1$

It is enough to know now that we can also use StatKey to calculate and visually see the critical values. Go to the “theoretical distributions” menu in StatKey at www.lock5stat.com and click on “t”. Put in 800 degrees of freedom and click “two-tail”. Since we are using a 5% significance level in two tails, each tail should have a proportion of 0.025 (2.5%). Notice StatKey gives the same critical values that Statcato gave.



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

P-value and Conclusions

Let us see if we can finish this test about the level of statistics student and the amount of alcoholic beverages consumed per week. The Statcato printout indicated that the P-value is 0.0653. Since the P-value is higher than the 5% significance level, we will fail to reject the null hypothesis. The claim was the null hypothesis so our conclusion should be that we do not have significant evidence to reject the claim.

Conclusion: There is not significant evidence to reject the claim that the level of stat student is not related to the amount of alcohol beverages per week.

Hypothesis Test - Two population means: confidence level = 0.95

Samples of population 1 in Math 140 alcohol...

Samples of population 2 in Math 075 alcohol...

	N	Mean	Stdev
Population 1	322	2.224	4.684
Population 2	481	1.470	6.884

Null hypothesis: $\mu_1 - \mu_2 = 0.0$

Alternative hypothesis: $\mu_1 - \mu_2 \neq 0.0$

* Population standard deviations are unknown. *

DOF = 800

Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.963, 1.963	1.846	0.0653

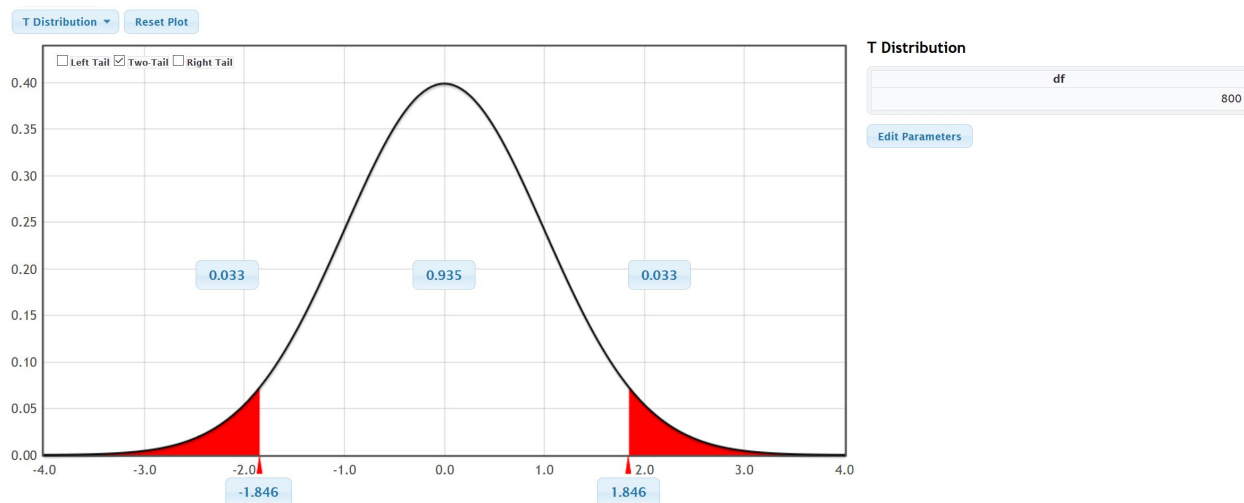
How was the P-value calculated?

P-values again can be calculated in different ways. A traditional approach would be to calculate the proportion in the tail or tails corresponding to the test statistic. Recall the degrees of freedom was 800. Using the theoretical distribution T calculator in StatKey, we can calculate the proportion in the tails. We entered the degrees of freedom and clicked "Two-Tail". We then entered our test statistic of +1.846 in the right tail since it was positive. The left tail automatically adapted. This was a two-tailed test, so we will need to add the proportions in the tails to get the P-value.

$$P\text{-value} = 0.033 + 0.033 = 0.066$$

P-value sentence: If the null hypothesis is true and the level of statistics student is not related to alcohol, then there is a 6.6% probability of getting this sample data or more extreme because of sampling variability.





We learned in the last chapter that we could also use randomized simulation to estimate the P-value. Open StatKey at www.lock5stat.com. When computing a two-population mean hypothesis test with StatKey, we will need the raw categorical and quantitative data. Open the "Math 075 140 combined survey Data Fall 2015" at www.matt-teachout.org. Copy the student level data and the alcoholic beverages data next to each other in a fresh excel spreadsheet. Under the "Randomization Hypothesis Tests" menu click on "Test for Difference in Means". Click on "Edit Data" and copy and paste both columns into StatKey. Click "Generate 1000 Samples" a few times. Remember this was a two-tailed test. The sample difference between the population 1 and population 2 was 0.754. Enter the sample difference of 0.754 into the right tail. Add the proportions in the tails to get the approximate P-value of $0.048 + 0.048 = 0.096$ or 9.6%.

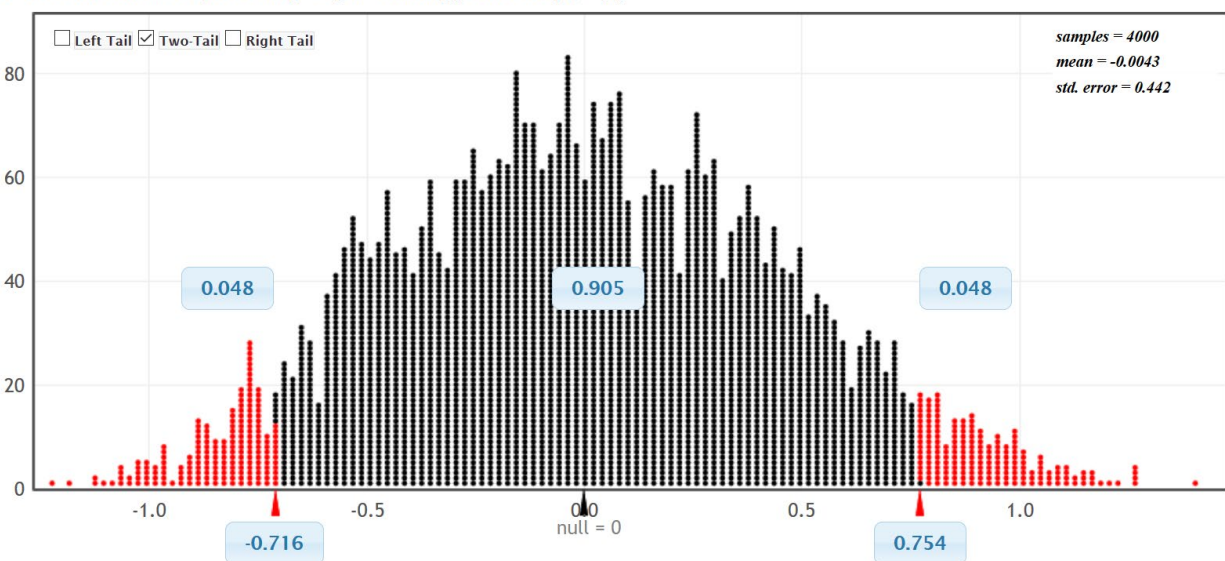
StatKey Randomization Test for a Difference in Means

Custom Dataset ▾ Show Data Table Edit Data Upload File Change Column(s)

Randomization method Reallocate Groups ▾

Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

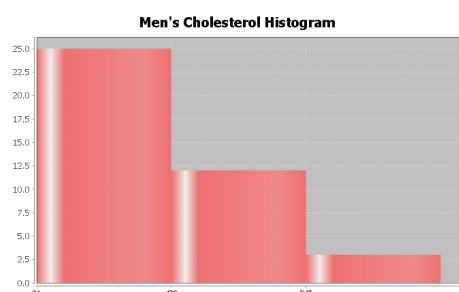
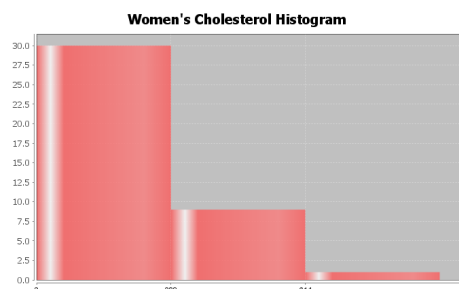
Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Example 2 (Two-population mean average T-test with Independent groups)

Some people believe that the population mean average cholesterol for men and women is the same, while others think that men's cholesterol is higher. Use the randomly collected health data at www.matt-teachout.org to test the claim that the mean average cholesterol for men (μ_1) is higher than the mean average cholesterol for women (μ_2). This claim would indicate that gender is related to cholesterol. Use the following Statcato printout, graphs and a 10% significance level.



Hypothesis Test: 2-Population Means

Help
×

F1

Inputs

☐ Samples in one column

Labels in column:

Values in column:

☒ Samples in two columns

Population 1:

Population 2:

☐ Summarized sample data

	Sample Size	Mean	Standard Deviation
Population 1:	<input type="text"/>	<input type="text"/>	<input type="text"/>
Population 2:	<input type="text"/>	<input type="text"/>	<input type="text"/>

Population Standard Deviations/Variances

☐ Population standard deviations known

σ_1 :

σ_2 :

☐ Assume population variances are equal

Alternative Hypothesis

Alternative Hypothesis:

Hypothesized Mean Difference:

Significance

☒ Significance Level: 0 - 1.00 (e.g. 0.05)

☐ Confidence Level: 0 - 1.00 (e.g. 0.95)



Hypothesis Test - Two population means: confidence level = 0.90

Samples of population 1 in C22 Men Chol

Samples of population 2 in C8 Women Chol

	N	Mean	Stdev
Population 1	40	395.225	292.412
Population 2	40	240.875	185.982

Null hypothesis: $\mu_1 - \mu_2 = 0.0$ Alternative hypothesis: $\mu_1 - \mu_2 > 0.0$

* Population standard deviations are unknown. *

DOF = 66

Significance Level	Critical Value	Test Statistic t	p-Value
0.10	1.295	2.817	0.0032

Null and alternative hypothesis

 $H_0 : \mu_1 = \mu_2$ (Gender and Cholesterol are NOT related) $H_A : \mu_1 > \mu_2$ (Gender and Cholesterol ARE related) (Claim)

Type of hypothesis test? Two-population Mean T-test (right tail with independent groups)

Assumptions? The data did pass all of the assumptions, so we can proceed with the hypothesis test.

- Both samples were collected randomly.
- Both samples pass the at least 30 or normal requirement. Even though both data sets were skewed right, they both had a sample size of 40.
- Data values within the samples were likely to be independent. These are simple random samples out of large populations. It is unlikely that the individual men in the data will be related. It is also unlikely that individual women will be related.
- Data values between the samples were likely to be independent. The groups were not matched pairs. They were not the same people measured twice or some other one to one pairing. Since the men and women were collected randomly out of a large population, it is unlikely they will be related.

T-test statistic = 2.817

Test Stat Sentence: The sample mean cholesterol for the men (395.225 mg/dL) is 2.817 standard errors above the sample mean cholesterol for the women (240.875 mg/dL).

- The right tail starts at the critical value of 1.295, so the test statistic definitely falls in the right tail and is significant.
- This tells us that the sample mean cholesterol for the men is significantly higher than for the women.
- The sample data significantly disagrees with the null hypothesis.

P-value = 0.0032 = 0.32%

P-Value Sentence: If H_0 is true, and men and women have the same population mean average cholesterol, then we had a 0.32% probability of getting the sample data or more extreme because of sampling variability.

- This is a low P-value. (The P-value of 0.32% is much smaller than the 10% significance level.)
- If H_0 is true, this tells us that the sample data was unlikely to have happened by random chance (sampling variability).
- A low P-value also indicates significance. This tells us that the sample mean cholesterol for the men is significantly higher than for the women.
- There is a significant disagreement between the sample data and the null hypothesis.



This chapter is from *Introduction to Statistics for Community College Students*,
 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
 under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Reject H_0 or Fail to reject H_0 ? Reject H_0 since the P-value (0.32%) is smaller than the 10% significance level.

Conclusion?

There is significant evidence to support the claim that the population mean average cholesterol for men is higher than the population mean average cholesterol for women. This also gives evidence that a gender is related to cholesterol.

(The random sample data significantly agrees with the claim that the population mean average cholesterol for men is higher than the population mean average cholesterol for women. We have a low P-value as evidence.)

Practice Problems Section 4A

(#1-10) Use each of the following two-population mean T-test statistics and the corresponding critical values to fill out the table.

	Type of Test	T-test stat	Sentence to explain T-test statistic.	Critical Value	Does the T-test statistic fall in a tail determined by a critical value? (Yes or No)	Are the sample means from the two groups significantly different or not? Explain.	Does sample data significantly disagree with H_0 ? Explain.
1.	Right Tailed	+1.383		+2.447			
2.	Left Tailed	-2.851		-1.773			
3.	Two Tailed	-1.501		± 2.006			
4.	Right Tailed	+3.561		+1.692			
5.	Two Tailed	+0.887		± 1.943			
6.	Left Tailed	-1.003		-2.759			
7.	Two Tailed	-4.416		± 1.994			
8.	Right Tailed	+0.275		+1.839			
9.	Left Tailed	-1.461		-1.674			
10.	Two Tailed	+2.330		± 2.138			

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by sampling variability or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.0007			10%			
12.	0.421			1%			
13.	8.71×10^{-5}			5%			
14.	0.339			1%			
15.	0.076			5%			
16.	0			10%			
17.	0.528			5%			



18.	0.0277			10%			
19.	3.04×10^{-6}			1%			
20.	0.178			5%			

21. Explain the difference between matched pair data and independent groups.
22. Explain the difference between random samples and random assignment.
23. List the assumptions that we need to check for a two-population mean hypothesis test from independent groups.
24. List the assumptions that we need to check for a two-population mean hypothesis test from matched pairs.
25. List the assumptions that we need to check for a two-population mean hypothesis test that is using experimental design.
26. Explain how to use a two-population mean hypothesis test to show that categorical and quantitative data are related.
27. Explain how to use a two-population mean hypothesis test to show there is a cause and effect between categorical and quantitative data.

(#28-33) *Directions:*

- a) *Determine if the following two-population mean tests are matched pair or independent groups*
- b) *Write the null and alternative hypothesis. Include relationship implications.*
- c) *Check all of the assumptions for a two-population mean T-test. Explain your answers. Does the problem meets all the assumptions?*
- d) *Write a sentence to explain the T-test statistic.*
- e) *Use the test statistics and the critical value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.*
- f) *Write a sentence to explain the P-value.*
- g) *Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.*
- h) *Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.*
- i) *Write a conclusion for the hypothesis test. Explain your conclusion in plain language.*
- j) *Is the categorical variable related to the quantitative variable? Explain your answer.*

28. The ACT exam is used by many colleges to test the readiness of high school students for college. Many high school students are now taking ACT prep classes. A local high school offers an ACT prep class, but wants to know if it really helps. Twenty students were randomly selected. They took the ACT exam before and after taking the ACT prep class. For each student the difference between the after and before scores were measured ($d = \text{after} - \text{before}$). Population 1 was the after prep class scores and population 2 was the before prep class scores. The mean of the differences was 1.5 ACT points with a standard deviation of 2.3 ACT points. A histogram of the differences yielded a bell shaped normal distribution. Use a 5% significance level to test the claim that the after prep class scores are higher than the before prep class scores. What does this data indicate about the relationship between taking a prep class or not and ACT scores.



N	Sample Mean	Stdev	Significance Level	Critical Value	Test Statistic t	p-Value
20	1.5	2.3	0.05	1.729	2.917	0.0044

29. A random sample of 20 male German Shepherds found that their average weight was 112 pounds with a standard deviation of 28 pounds. A random sample of 14 male Dobermans found that their average weight is 107 pounds with a standard deviation of 24 pounds. Assume that weights are normally distributed. Use the Statcato printout below and a 5% significance level to test the claim that the population mean average weight of male German Shepherds (population 1) is more than the population mean average weight of male Doberman Pinchers (population 2). What does this data indicate about the relationship between the weight and the type of dog?

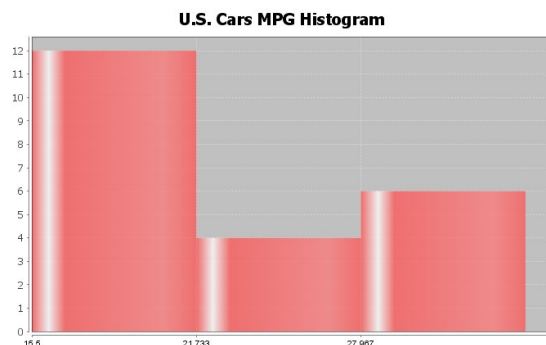
	N	Mean	Stdev
German Shep Sample 1	20	112.0	28.0
Doberman Sample 2	14	107.0	24.0

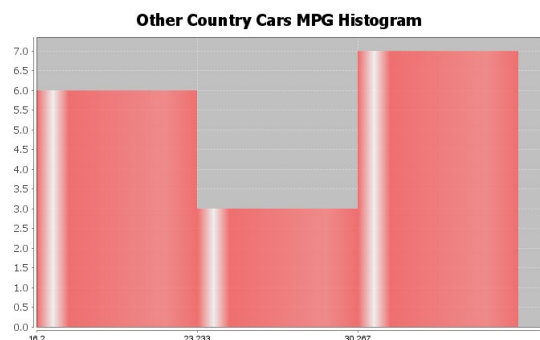
Significance Level	Critical Value	Test Statistic t	p-Value
0.05	1.697	0.558	0.2906

30. Cotinine is an alkaloid found in tobacco and is used as a biomarker for exposure to cigarette smoke. It is especially useful in examining a person's exposure to second hand smoke. A random sample of 32 non-smoking American adults was collected. These adults were not smokers and did not live with any smokers. The average cotinine level for this sample was 7.2 ng/mL with a standard deviation of 5.8 ng/mL. A second random sample of 35 non-smoking American adults was then collected. These adults did not smoke themselves, but did live with one or more smokers. The mean average cotinine level for this sample was 28.5 ng/mL and had a standard deviation of 11.4 ng/mL. Use a 1% significance level to test the claim that people that do not live with smokers have a lower cotinine level than those people that do live with smokers. What does this data indicate about the relationship between cotinine levels and living with a smoker or not.

Significance Level	Critical Value	Test Statistic t	p-Value
0.01	-2.402	-9.758	$1.4751 \cdot 10^{-13}$

31. We want to see if the country a car is made in is related to its gas mileage in miles per gallon. Specifically we wanted to see if cars made in the U.S. have a lower population mean average mpg than those made outside the U.S. We used the random car data at www.matt-teachout.org and a 5% significance level to create the following graphs and statistics with Statcato. Check the assumptions and perform the hypothesis test.





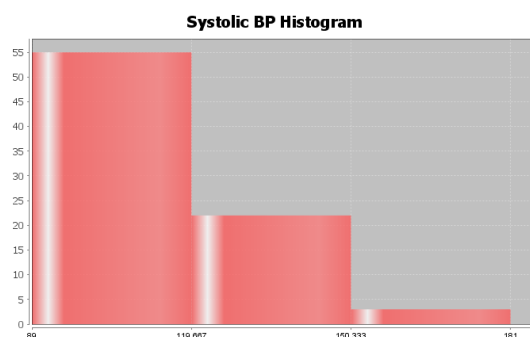
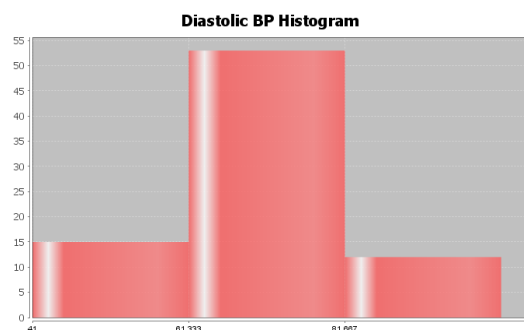
	N	Mean	Stdev
Population 1 USA mpg	22	22.995	6.054
Population 2 Other Country mpg	16	27.188	6.601

* Population standard deviations are unknown. *

DOF = 30

Significance Level	Critical Value	Test Statistic t	p-Value
0.05	-1.697	-2.001	0.0273

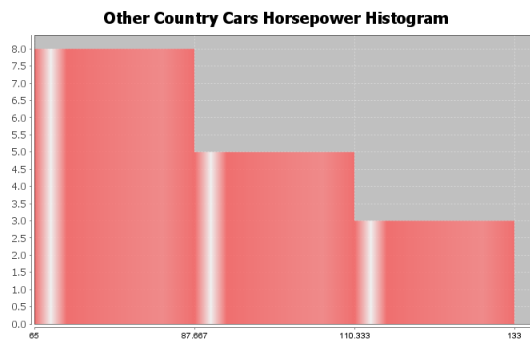
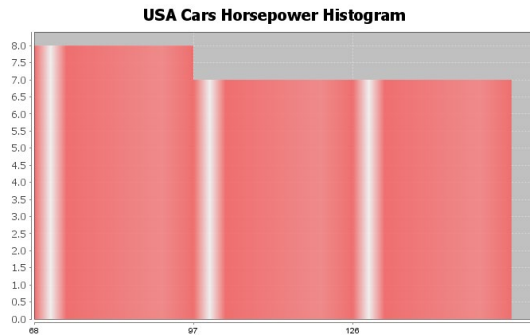
32. We want to test the claim that the diastolic blood pressure of a person is less than the systolic blood pressure of a person. We used the random health data at www.matt-teachout.org, Statcato, and a 1% significance level to create the following graphs and statistics. Check the assumptions and perform the hypothesis test. Notice that since the diastolic and systolic blood pressures came from the same randomly selected adults, we used a matched pair calculation.



This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

N	Sample Mean	Stdev	Significance Level	Critical Value	Test Statistic t	p-Value
80	-44.525	10.077	0.01	-2.375	-39.521	$4.188 \cdot 10^{-54}$

33. We want to see if the country a car is made in is related to the horsepower of the car. Specifically we wanted to see if cars made in the U.S. have a higher population mean average horsepower than those made outside the U.S. We used the random car data at www.matt-teachout.org and a 10% significance level to create the following graphs and statistics with Statcato. Check the assumptions and perform the hypothesis test.



	N	Mean	Stdev
Population 1 USA car horsepower	22	110.182	26.383
Population 2 Other Country car horsepower	16	90.125	22.408

Significance Level	Critical Value	Test Statistic t	p-Value
0.10	1.306	2.526	0.0081



(#34-37) Directions: For each problem, answer the following questions.

- Determine if the following two-population mean tests are matched pair or independent groups
- Write the null and alternative hypothesis. Include relationship implications.
- Use randomized simulation to calculate the P-value. Write a sentence to explain the P-value.
- Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.
- Use the sample mean difference and the standard error in the simulation to calculate the T-test statistic.

$$T = \frac{\text{Sample Mean Difference}}{\text{Standard Error}}$$

- Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- Write a conclusion for the hypothesis test. Explain your conclusion in plain language.
- Is the categorical variable related to the quantitative variable? Explain your answer.

34. Go to StatKey at www.lock5stat.com. Under the “Randomization Hypothesis Tests” menu, click on “Test for Difference for Difference in Means”. Under the data sets menu on the top left, click on “Commute Time (Atlanta vs St. Louis)”. This took a random sample of people from Atlanta (population 1) and a random sample of people from St. Louis (population 2). Use randomized simulation and a 5% significance level to test the claim that the mean average commute time for people in Atlanta is greater than the mean average commute time for people from St. Louis. What does this data indicate about the relationship between the city and the commute time?

35. Go to StatKey at www.lock5stat.com. Under the “Randomization Hypothesis Tests” menu, click on “Test for Single Mean”. Under the data sets menu on the top left, click on “Pulse Rate Difference (Quiz – Lecture)”. An experiment was done on college students to determine if heart rate is related to taking a quiz or not. The heart rates of students were measured on a day they were taking a quiz (population 1) and again on a day when there was just lecture (population 2). The same students were measured twice. Use randomized simulation and a 1% significance level to test the claim that the mean average heart rate difference between the quiz and lecture days is greater than zero. This will indicate that the heart rates on quiz days tend to be higher than lecture days. What does this data indicate about the relationship between the heart rate and taking a quiz or not?

36. Go to StatKey at www.lock5stat.com. Under the “Randomization Hypothesis Tests” menu, click on “Test for Difference for Difference in Means”. Under the data sets menu on the top left, click on “Exercise Hours (Male vs Female)”. This took a random sample of male adults (population 1) and a random sample of female adults (population 2). Use randomized simulation and a 10% significance level to test the claim that the mean average amount of time that males and females exercise is the same. What does this data indicate about the relationship between exercise hours and gender?

37. Use StatKey and the random health data to test the claim that the population mean average pulse rate for women is higher than for men. Go to www.matt-teachout.org and click on the statistics tab and then the data sets tab. Open the health data. Copy and paste the gender data and pulse data columns next to each other in a fresh excel spreadsheet. Now copy the two columns. Go to StatKey at www.lock5stat.com. Under the “Randomization Hypothesis Tests” menu, click on “Test for Difference for Difference in Means”. Under “Edit Data”, paste the gender and pulse rate columns into StatKey. Click on “Generate 1000 Samples” a few times. Click on “Right Tail” and put in the sample mean difference of 6.9 beats per minute in the bottom box in order to estimate the P-value. Now answer the questions above.



Section 4B – Categorical/Quantitative Relationships: ANOVA

Introduction

In the last section, we saw that we could use a two-population mean average hypothesis test to determine if categorical and quantitative variables are related or not. If the mean averages were the same in two groups that would indicate that the categorical variable that determines the groups is not related to the quantitative variable mean average.

$H_0 : \mu_1 = \mu_2$ (categorical variable is not related to the quantitative variable)

$H_A : \mu_1 \neq \mu_2$ (categorical variable is related to the quantitative variable)

Many times, categorical data has more than just two options. This would mean that we would need to compare three or more population means.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_k$ (categorical variable is not related to the quantitative variable)

$H_A : \text{at least one population mean is } \neq$ (categorical variable is related to the quantitative variable)

Unfortunately, a T test statistic can only compare two things at a time and cannot handle a hypothesis test involving three or more groups. To compare three or more population means, we will need to use ANOVA.

ANOVA

ANOVA stands for “Analysis of Variance”. If you remember, variance is the square of the standard deviation. Variance measures the variability from the mean. There are two specific variances that are compared in an ANOVA test, the variance between the groups and the variance within the groups. The variance between the groups is a measure of how different the groups are. It measures how much variability each of the sample means are from the mean of all the groups combined. The variance within the groups measures the amount of variability that data values in each group are from their own sample mean. In ANOVA, we compare the variance between the groups to the variance within the groups.

ANOVA tests use the F-test statistic. In ANOVA, the F-test statistic divides the variance between the groups to the variance within the groups.

F Test Statistic Sentence: The ratio of the variance between the groups to the variance within the groups.

Calculating and Interpreting the F-test Statistic

As with all difficult calculations in statistics, use a computer program to calculate the F-test statistic. Never calculate it by hand. Always focus more on interpretation than on calculation. Let us see if we can better understand how the F-test statistic works.

Variance divides the sum of squares of the differences by the degrees of freedom.

$$\text{Variance} = \frac{\text{Sum of Squares}}{\text{Degrees of Freedom}}$$

To calculate the variance between the groups, the computer calculates the sum of squares between the groups and then divides by the degrees of freedom. The sum of squares between the groups subtracts the mean of all the groups combined (\bar{x}) from the sample means (\bar{x}_i) for each group. It squares the differences and adds them. Since the variance between calculations is based on the number of groups, the degrees of freedom between is the number of groups – 1 or “k – 1”.

$$\text{Variance} = \frac{\text{Sum of Squares Between}}{\text{Degrees of Freedom Between}} = \frac{\sum (\bar{x}_i - \bar{x})^2}{k-1}$$



To calculate the variance within the groups, we divide the sum of squares within each group divided by the sum of squares within. The “sum of squares within” subtracts each data value minus its own sample mean, squares the differences and adds them up. If we look at the degrees of freedom for each data set ($n_i - 1$) and add them up for each group, we will get the “degrees of freedom within”.

$$\text{Variance} = \frac{\text{Sum of Squares Within}}{\text{Degrees of Freedom Within}} = \frac{\sum (x - \bar{x}_i)^2}{\sum (n_i - 1)}$$

There is a beauty in the mathematics behind the F test statistic. The total number of data values for all of the groups combined minus one is often called the total degrees of freedom. There is also a total sum of squares.

Sum of Squares Between + Sum of Squares Within = Total Sum of Squares

Degrees of Freedom Between + Degrees of Freedom Within = Total Degrees of Freedom

Variance Between + Variance Within = Total Variance

As we said, the F test statistic divides the Variance between the groups by the Variance within the groups. We often say that the F-test statistic is the ratio of two variances. In ANOVA, it is the ratio of the variance between the groups to the variance within the groups.

$$\text{F test statistic} = \frac{\text{Variance Between the Groups}}{\text{Variance Within the Groups}} = \frac{\left(\frac{\text{Sum of Squares Between}}{\text{Degrees of Freedom Between}} \right)}{\left(\frac{\text{Sum of Squares Within}}{\text{Degrees of Freedom Within}} \right)}$$

Computer Programs will often give you sum of squares, degrees of freedom, and variances for the F test statistic. Look at the following printout. This test used a 5% significance level.

Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	4	10484529.98982	2621132.49746	7.92175	2.4248	$7.03917 \cdot 10^{-6}$
Error (Within Groups)	170	56249274.83547	330878.08727			
Total	174	66733804.82529				

Let us see if we understand what we are seeing. Notice the MS (mean sum of squares) is the sum of squares (SS) divided by degrees of freedom (df).

“MS Treatment (Between Groups) is the variance between the groups 2621124.8 which was calculated by dividing the sum of squares (SS Between) 10484529.98982 by the degrees of freedom (DOF Between) 4.

“MS Error (Within Groups) is the variance within the groups 330878.08727 which was calculated by dividing the sum of squares (SS Within) 56249274.83547 by the degrees of freedom (DOF Within) 170.

So the F-test statistic is calculated by dividing the variances (MS).

$$\text{F test statistic} = \frac{\text{Variance Between the Groups}}{\text{Variance Within the Groups}} = \frac{2621124.8}{330878.08727} = 7.92175$$

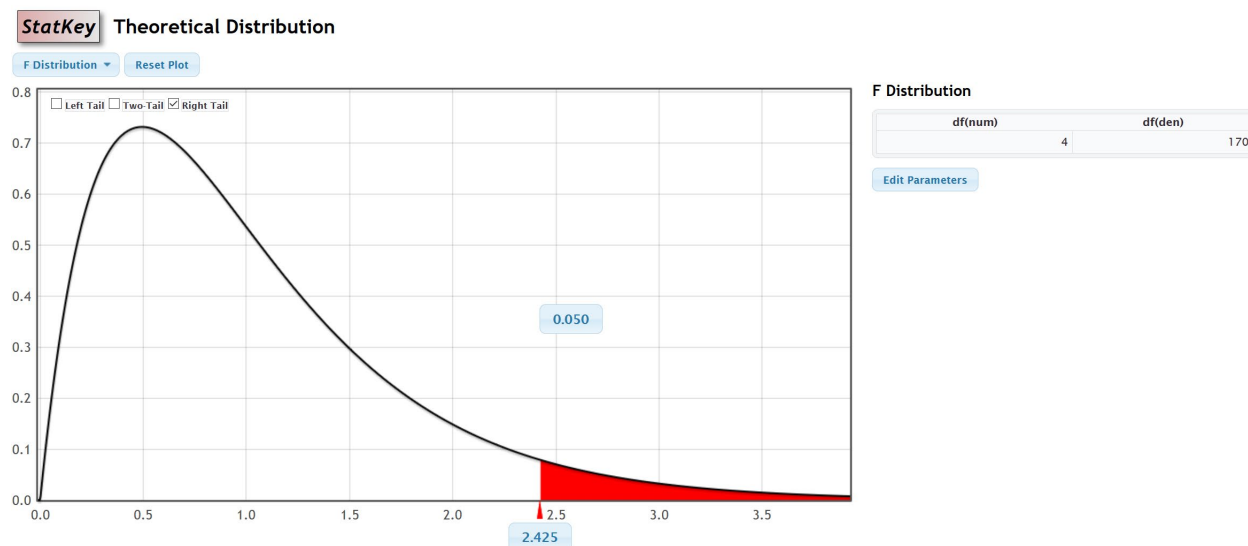
So the variance between the groups is almost 8 times greater than the variance within the groups.

Is this F test statistic significant?



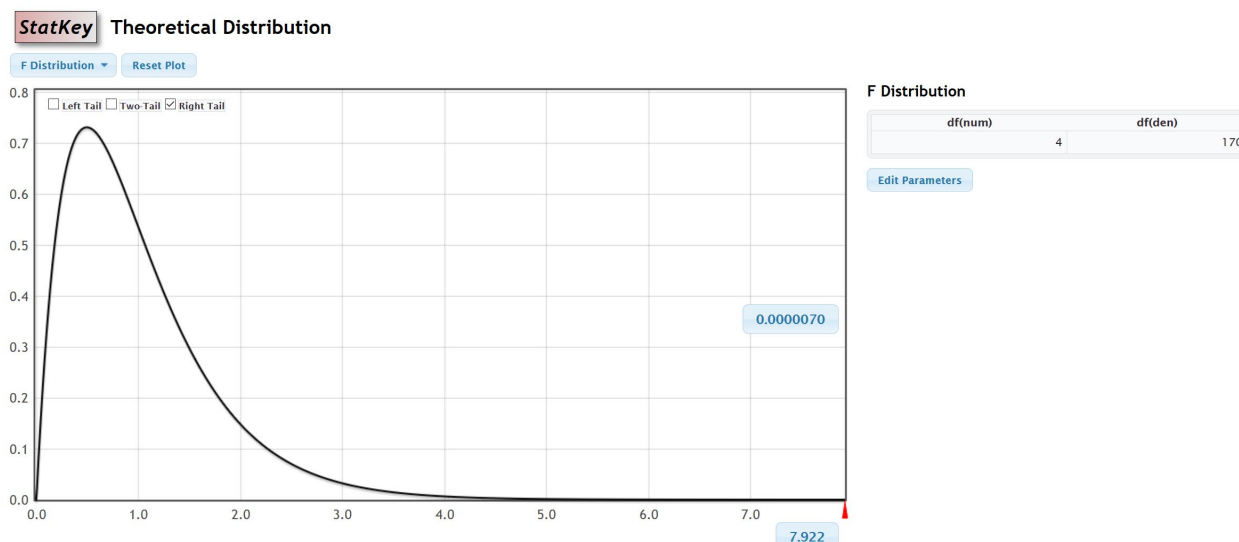
Notice Statcato has calculated a critical value to compare the test statistic to. ANOVA is always a right tailed test. Remember the test statistic needs to be in the tail determined by the critical value in order to be significant. Statcato thinks that the F test statistic has to be 2.4248 or higher to be significant. Our test statistic is 7.9217, which is definitely larger than the critical value 2.4248 and falls in the right tail. So our F test statistic is significant. Therefore, the F test statistic is significantly large and the variance between the groups is significantly greater than the variance within the groups. As with all test statistics, this also tells us that the sample data significantly disagrees with the null hypothesis.

We could have looked up the critical value with StatKey as well. You will need to know the degrees of freedom between (numerator degrees of freedom) which was four and the degrees of freedom within (denominator degrees of freedom). Go to www.lock5stat.com and open StatKey. Under “Theoretical Distributions” click on “F”. Put in the numerator degrees of freedom as 4 and the denominator degrees of freedom as 170. Since ANOVA is always a right tailed test and the significance level is 5%, simply click on “Right Tail” and enter 0.05 for the right tail proportion. Notice the number on the bottom is the critical value. It is about the same as what Statcato gave.



Notice we can also use the same F theoretical curve to calculate the P-value with StatKey. Remember the numerator degrees of freedom is 4 and the denominator degrees of freedom is 170. Now just put the F test statistic in the bottom box in the right tail. The proportion is the P-value. Notice the P-value calculated by StatKey is very close to the P-value calculated by Statcato.





Notes about the F-test statistic

- In a fraction, when the numerator is significantly larger than the denominator, the overall fraction is large. If the variance between the groups is much larger than the variance within the groups, this will give a large F-test statistic (and a small P-value) and indicates that the sample means are significantly different. A small P-value indicates that the sample data is unlikely to happen by random chance. We will reject the null hypothesis that the population means are the same. We are also rejecting that the categorical and quantitative variables are not related and supporting the alternative hypothesis that the variables are related.
- In a fraction, when the numerator is the same or smaller than the denominator, the overall fraction is small. So if the variance between the groups is much smaller than the variance within the groups, this will give a small F-test statistic (and a large P-value) and indicates that the sample means are not significantly different. A large P-value indicates that the sample data could have happened by random chance. We will fail to reject the null hypothesis that the population means are the same. In other words, the population means might be the same and the categorical and quantitative variables are probably not related.
- The F-test statistic can also be used in a two-population variance or two-population standard deviation hypothesis test. In that case, it compares the variance from two populations.



Here is the summary table from last chapter to remind you of the key decisions in a hypothesis test.

	Significant Test Statistic	Test Statistic NOT Significant
	<i>(Test Statistic falls in tail determined by the critical value or values)</i>	<i>(Test Statistic does NOT fall in tail determined by the critical value or values)</i>
	OR	OR
	Small P-value	Large P-value
	<i>(P-value \leq significance level)</i>	<i>(P-value $>$ significance level)</i>
	OR	OR
	Sample Data in Tail	Sample Data NOT in Tail
	<i>(when simulating the Null Hypothesis)</i>	<i>(when simulating the Null Hypothesis)</i>
Is the sample data significantly different than H_0?	Yes. Significantly different	Not Significantly different
Could the sample data happen by random chance (sampling variability) if H_0 is true?	Unlikely	Could happen
Reject H_0 or Fail to Reject H_0?	Reject H_0	Fail to Reject H_0
Is there significant Evidence?	Yes. Is evidence	No evidence

Assumptions for an ANOVA hypothesis test

- The quantitative samples should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- Data values between the samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape
- No standard deviation for any sample is more than twice as large as any other sample.

Notice that we must have a random or representative sample. As with all mean average hypothesis tests, we require the sample size to be at least 30 or have a normal shape. Data values within the samples and between the samples should be independent of each other. This again is a difficult assumption to assess. If we have a small simple random sample out of a very large population, then the data values are unlikely to be related. ANOVA is based on variance, so variability in the samples is very important. If one sample has a lot more variability than the others do, this can be a problem. Therefore, we want all of our sample standard deviations to be close. An often-used rule is that no sample standard deviation can be more than twice as large as any other can. Notice that to check these assumptions, we need to look at the sample sizes, sample means and sample standard deviations for each of our groups. We should also look at the shape of the samples with histograms or dot plots.



ANOVA Example 1: Mean Average Salaries for people living in five states in Australia.

Suppose we want to compare the mean average weekly salary for people living in five states in Australia. The states are Northern Territory, New South Wales, Queensland, Victoria, and Tasmania. We claim that the mean average salary of people is related to where they live. To support this claim, we will need to show that the mean average salaries are different in these states. As with all multiple population hypothesis tests, you should label the populations. To perform this test, adults were randomly selected from each of the five states. We will be using a 5% significance level.

μ_1 : Northern Territory

μ_2 : New South Wales

μ_3 : Queensland

μ_4 : Victoria

μ_5 : Tasmania

Here is the null and alternative hypothesis for the ANOVA test. Remember an ANOVA is a multiple μ test for three or more groups. Notice that if the population mean average salary is the same, then it does not matter which state the person lives in. This implies that the state (categorical variable) is not related to the salary (quantitative variable). If at least one population mean average salary is different, then it does matter which state the person lives in. This implies that the state (categorical variable) is related to the salary (quantitative variable). Again, we see that “not related” is the null hypothesis and “related” is the alternative.

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (*states in Australia are not related to salary*)

Ha: at least one is \neq (CLAIM) (*states in Australia are related to salary*)

When doing an ANOVA test, it is good to find the sample size (n), the sample mean of each group, and the standard deviation for each group. We also will need to create histograms to check the shape of our samples. Go to www.matt-teachout.org and click on the “statistics” tab and then “data sets”. You can either open the Australia Salary data Statcato file in Statcato or open the excel file and copy and paste it into Statcato. The adults in this sample data were randomly selected. To calculate the sample sizes, means and standard deviations, go to the “statistics” menu in Statcato, then click on “basic statistics” then “descriptive statistics”. To create histograms, go the “graph” menu and click on histogram. In small data sets like these, I prefer three bins (bars). It makes it easier to see the shape. In addition, if you click on “Show Legend” the computer will also make a title for the graph. Here is the sample statistics and graphs from Statcato.



Descriptive Statistics

✕

Help

F1

Inputs

Input Variable(s):
c1-c5

Enter valid column names separated by space. For a continuous range of columns, separate using dash (e.g. C1-C30).

By Variable (optional):
▼

Results

Store Results in:
☐ New datasheet

Statistics

☐ Select all statistics

☒ Mean
☐ SE of mean
☒ Standard deviation
☐ Variance
☐ Coefficient of variation

☐ Trimmed mean: cutoff % % of values to be trimmed (between 0 and 100)
☐ Sum
☐ Minimum
☐ Maximum
☐ Range

☐ First quartile
☐ Median
☐ Third quartile
☐ Interquartile range
☐ Mode
☐ Percentile:
e.g. 10 for the 10th percentile

☐ N nonmissing
☐ N missing
☒ N total
☐ Cumulative N
☐ Percent
☐ Cumulative Percent

☐ Sum of squares
☐ Skewness
☐ Kurtosis
☐ MSSD

OK Cancel

Descriptive Statistics

Variable	Mean	Standard Deviation
C1 North Territory	1534.540	701.525
C2 New South Wales	1536.823	677.140
C3 Queensland	1368.291	536.319
C4 Victoria	1149.050	516.553
C5 Tasmania	898.695	386.354

Variable	N total
C1 North Territory	35
C2 New South Wales	35
C3 Queensland	35
C4 Victoria	35
C5 Tasmania	35

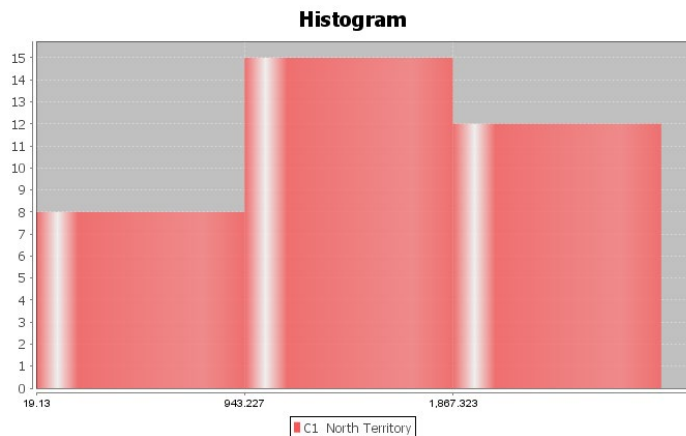


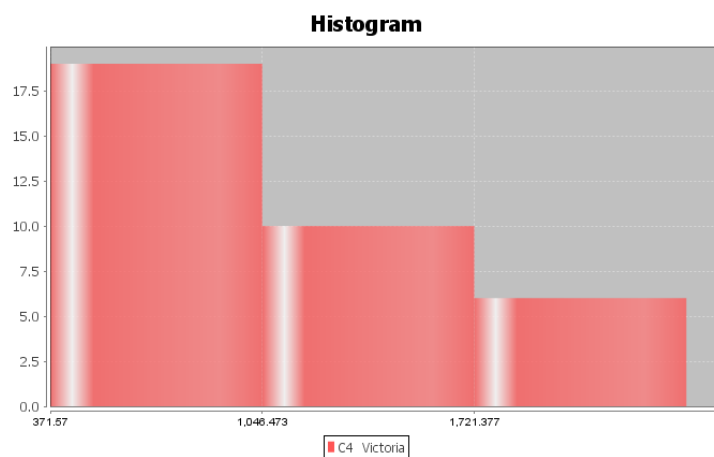
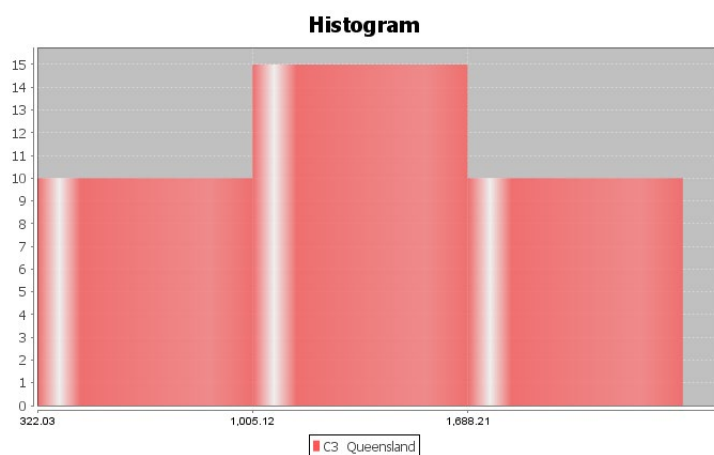
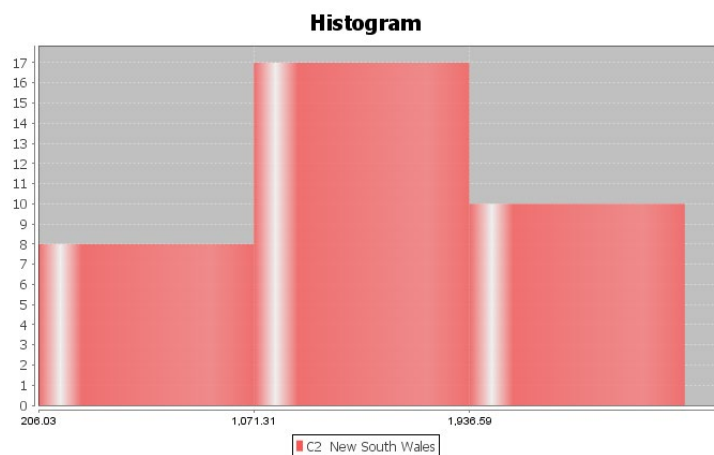
This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

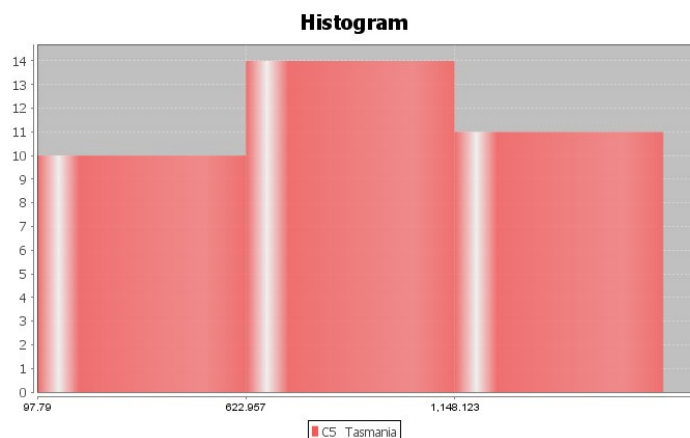
Histogram

Help

<p>Graph Variables</p> <p>Graph Variables:</p> <p>Ctrl-click to select multiple variables</p> <p>C1 North Territory C2 New South Wales</p> <p>Grouped By Categories in: [optional]</p> <p>Heights of bars represent:</p> <p><input checked="" type="radio"/> Frequency</p> <p><input type="radio"/> Relative Frequency</p>	<p>X-Axis</p> <p>X-axis (horizontal)</p> <p><input checked="" type="radio"/> Provide the number of classes, minimum, and maximum</p> <p>Class width = (maximum - minimum) / classes</p> <p>Number of bins (classes): 3</p> <p>Minimum: [] Maximum: [] [automatic if left blank]</p> <p><input type="radio"/> Provide the class width and the minimum</p> <p>Class width: []</p> <p>Minimum: [] [automatic if left blank]</p> <p>Label: []</p> <p>Position of tick marks: <input type="radio"/> Center of bar <input checked="" type="radio"/> Between bars</p>
<p>Other Options</p> <p>Plot</p> <p>Title: Histogram</p> <p><input checked="" type="checkbox"/> Show Legend</p>	<p>Y-Axis</p> <p>Y-axis (vertical)</p> <p>Label: []</p> <p>Tick mark units: [] automatic if left blank</p>
<p>OK Cancel</p>	







Assumptions: Notice that this data passes all of the assumptions for the ANOVA hypothesis test.

1. The sample data should be random or representative of the population. [Yes. The sample data sets were collected randomly.](#)
2. Each sample should have a sample size of at least 30 or be nearly normal. [Yes. All of the samples had a nearly normal shape except for the data from Victoria, which was skewed right. The sample size for all of the samples was 35. So even though data from Victoria was skewed right, its sample size was still over 30. All of the other samples sizes were over 30 and normal.](#)
3. Data values within the samples and between the samples should be independent of each other. [Yes. Since we are dealing with small random samples out of millions in the populations, it is unlikely that these data values are related.](#)
4. The sample standard deviations for the groups should be close. No standard deviation should be more than twice as large as any other should. [Yes. The sample standard deviations are close. No sample standard deviation is more than twice as large as any other sample standard deviation. Notice that the smallest standard deviation was 386.3 and the largest was 701.5 and all of the others are in between.](#)

Some data scientists like to create a side-by-side boxplot when performing an ANOVA test. This is surprising since the ANOVA test looks at means and standard deviations for center and spread, yet box plots look at the median and interquartile range (IQR). The boxplot can still show us general tendencies about shape, center and spread.

In Statcato, go to the “Graph” menu and click on “Box Plot”. Hold the control key down and highlight all five of the data sets. You can create a vertical or horizontal box plot. “Show Legend” will create a title. Do NOT click on the “Group By” button. Here is the box plot from Statcato.



Box Plot

Help

Graph Variables

Graph Variables:

Ctrl-click to select multiple variables

C3 Queensland

C4 Victoria

C5 Tasmania

<

>

Grouped By Categories in: [optional]

Graph Options

Plot Title:

X-axis Label:

Y-axis Label:

Orientation:

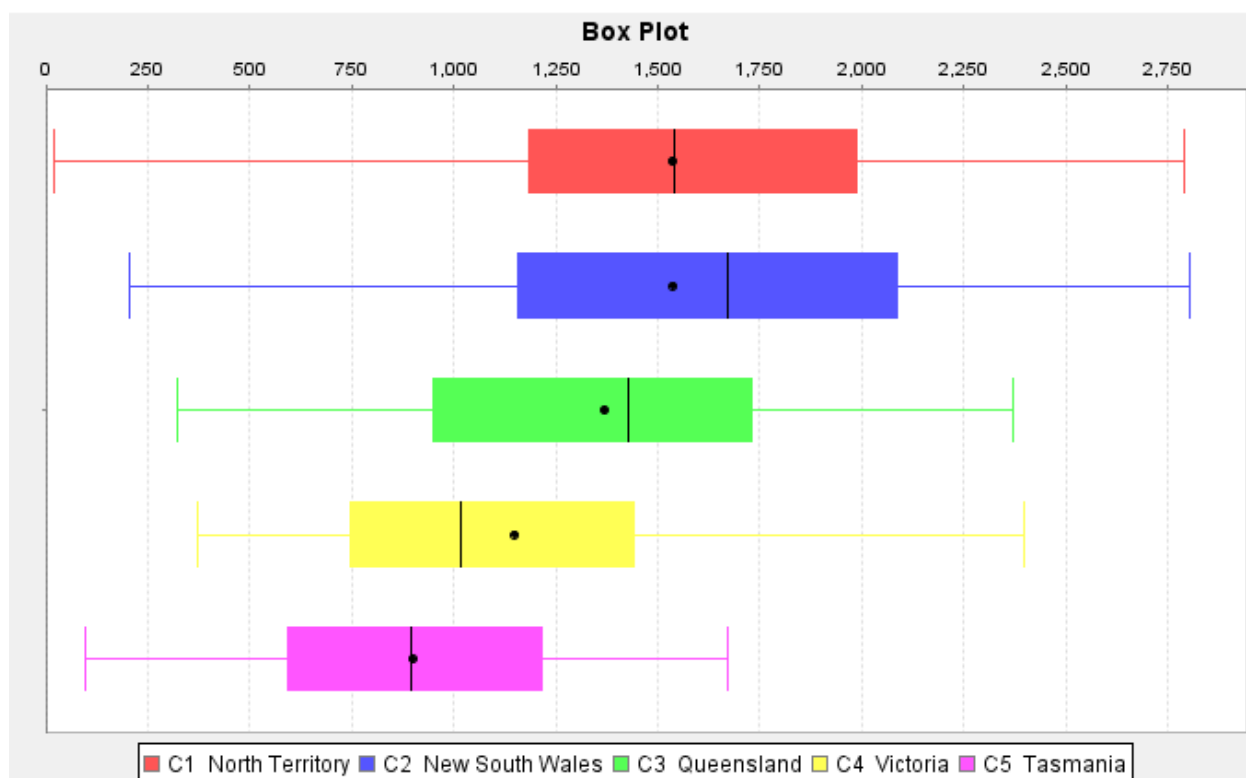
☒ Horizontal

☐ Vertical

☒ Show Legend

OK

Cancel



This chapter is from *Introduction to Statistics for Community College Students*,
 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
 under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

This graph tells a lot. We can see that the centers are quite different. The length of the box is a measure of spread (IQR). The lengths of the boxes are all pretty similar. If one box was more than twice as long as another was, this might indicate that one group has a lot more variability than another does. We can also get a sense of the shapes of these data sets, though separate histograms are better. So this side-by-side boxplot shows us that the variability is similar in the groups but the centers are quite different. Only the data from Victoria looks skewed right.

The key question: Are these sample means different because of sampling variability (random chance) OR are they different because at least one of the populations really is different?

To answer this, we need the F test statistic and a P-value.

How to do an ANOVA test with Statcato

Copy and paste your raw quantitative data from each group into some columns in Statcato.

To calculate the F-test statistic and P-value, go to the “statistics” menu, then “Analysis of Variance”, then “One-Way ANOVA”.

Statistics → Analysis of Variance → One-Way ANOVA

Hold the control key down to select the columns where your data is and push “add to list”. Select your significance level and push “OK”. Here is the printout we got. Notice this is the same printout we were looking at before.

One-way ANOVA: Significance level = 0.05

Selected column variables: C1 North Territory C2 New South Wales C3 Queensland C4 Victoria C5 Tasmania

Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	4	10484529.98982	2621132.49746	7.92175	2.4248	$7.03917 \cdot 10^{-6}$
Error (Within Groups)	170	56249274.83547	330878.08727			
Total	174	66733804.82529				

Let us see if we understand what we are seeing. Notice the MS (variance) is the sum of squares (SS) divided by degrees of freedom (DOF).

MS (Treatment) is the variance between the groups (2621124.8)

MS (Error) is the variance within the groups (330878.43)

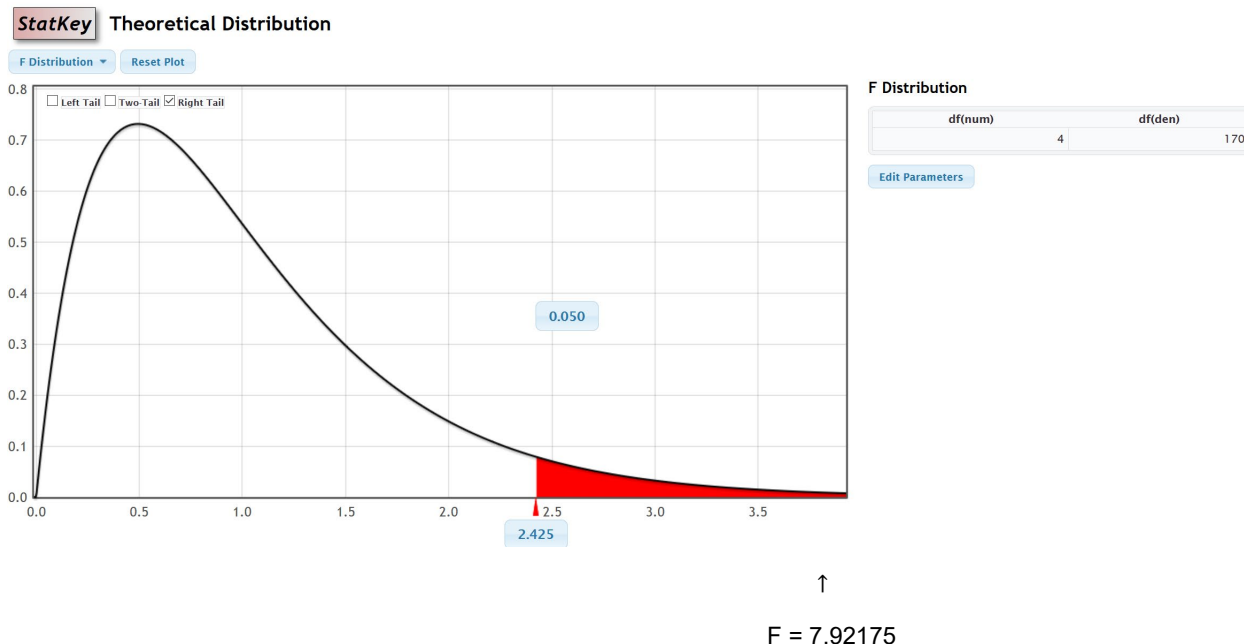
So the F-test statistic is calculated by the formula

$$F \text{ test statistic} = \frac{\text{Variance Between the Groups}}{\text{Variance Within the Groups}} = \frac{2621124.8}{330878.08727} = 7.92175$$

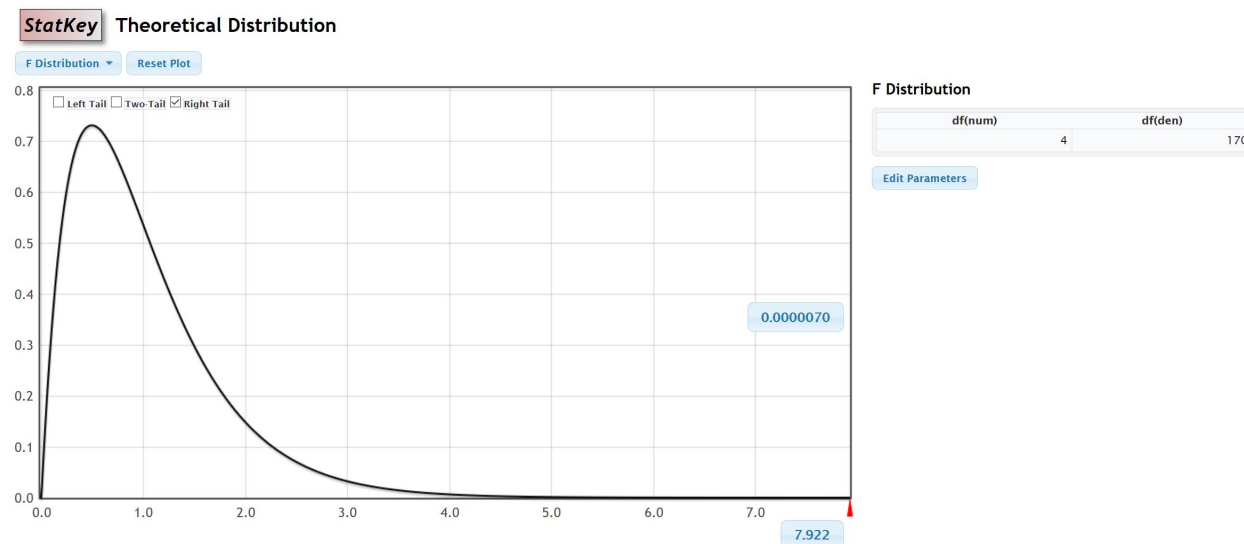
So the variance between the groups is almost 8 times greater than the variance within the groups. Is this significantly large for an F?

Notice Statcato has calculated a critical value to compare the test statistic to. Remember the test statistic needs to fall in the tail determined by the critical value to be significant. ANOVA is always a right tailed test. Look at the following picture created by StatKey. We see that our test statistic is 7.9217 falls in the right tail. So the F test statistic is significant. This also tells us that the sample data significantly disagrees with the null hypothesis and that the variance between the groups is significantly greater than the variance within the groups. Otherwise, the F test statistic would not have fallen in the right tail.





Notice we can also use the same F theoretical curve to calculate the P-value with StatKey. Remember the numerator degrees of freedom is 4 and the denominator degrees of freedom is 170. Now just put the F test statistic in the bottom box in the right tail. The proportion is the P-value. Notice the P-value calculated by StatKey is very close to the P-value calculated by Statcato.



The test statistic fell in the tail determined by the critical value, so the sample data does significantly disagree with the null hypothesis.

Notice that in our printout from Statcato, we got the following P-value: “7.039 x 10⁻⁶”. This is scientific notation. Move the decimal six places to the left to get the P-value as a decimal.

P-value = 0.00000704



This chapter is from *Introduction to Statistics for Community College Students*,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed
under a “CC-By” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

The actual P-value is very close to zero and much lower than a 5% significance level. From our study of P-values, we know this is very significant and unlikely to happen by random chance (sampling variability).

Since the P-value is less than our significance level, we should reject the null hypothesis.

Conclusion

Recall the claim was the alternative hypothesis that where a person lives is related to the salary. To show this we needed to have evidence to support that at least one state was different from the others (alternative hypothesis). Since we rejected the null, we support this claim. Our P-value is very small and our F test statistic very large, so we have significant evidence.

Conclusion: There is significant evidence to support the claim that the mean average salaries of people in Northern Territory, New South Wales, Queensland, Victoria, and Tasmania are different and that the state a person lives in is related to the salary.

Note: Remember “relationship” does not mean “causation” though. This was not an experiment and did not control confounding variables. There are many reasons why a persons’ salary is high or low. It would be wrong to say that the place a person lives causes their salary to be low or high.

Simulation

Remember we can also estimate the P-value and determine significance with randomized simulation. Go to www.lock5stat.com and open StatKey. We will need to go to www.matt-teachout.org and open the “Australia Salary Data” in Excel. In Statcato, we needed the quantitative data separated by group, but in StatKey, we need the raw categorical and quantitative data. StatKey will separate the data. In the excel spreadsheet you will see the column that says, “State in Australia” and “Salary”. Copy these two data sets together. Under the “More Advanced Randomization Tests” menu click on “ANOVA for Difference in Means”. Click on “Edit Data” and paste in the state and salary columns.



Edit data

State in Australia

Salary \$

North Territory

2034.68

North Territory

1228.05

North Territory

1504.05

North Territory

1975.87

North Territory

1542.29

North Territory

2338.33

North Territory

2368.36

North Territory

916.36

North Territory

1644.29

North Territory

1281.53

North Territory

1426.37

North Territory

1351.88

North Territory

2791.42

North Territory

1141.1

North Territory

2001.56

North Territory

1943.8

North Territory

1371.32

North Territory

1741.07

North Territory

1909.9

North Territory

1859.08

☒ Data has header row

Manually edit the values above or paste a tab or comma seperated file into the box and click Ok. The file must have only two columns where the first column is the categorical variable and the second is the quantitative.

Ok

Notice under “Original Sample”, StatKey has calculated the F-test statistic for you along with the sample means, sample sizes and sample standard deviations. If you wish to see the variance between and the variance within calculations click on “ANOVA Table”. It looks similar to the Statcato printout.



Original Sample

ANOVA Table

$$n = 175, F = 7.922$$

Statistics	North Territory	New South Wales	Queensland	Victoria	Tasmania	Overall
Sample Size	35	35	35	35	35	175
Mean	1534.5	1536.8	1368.3	1149.1	898.7	1297.5
Standard Deviation	701.5	677.1	536.3	516.6	386.4	619.3

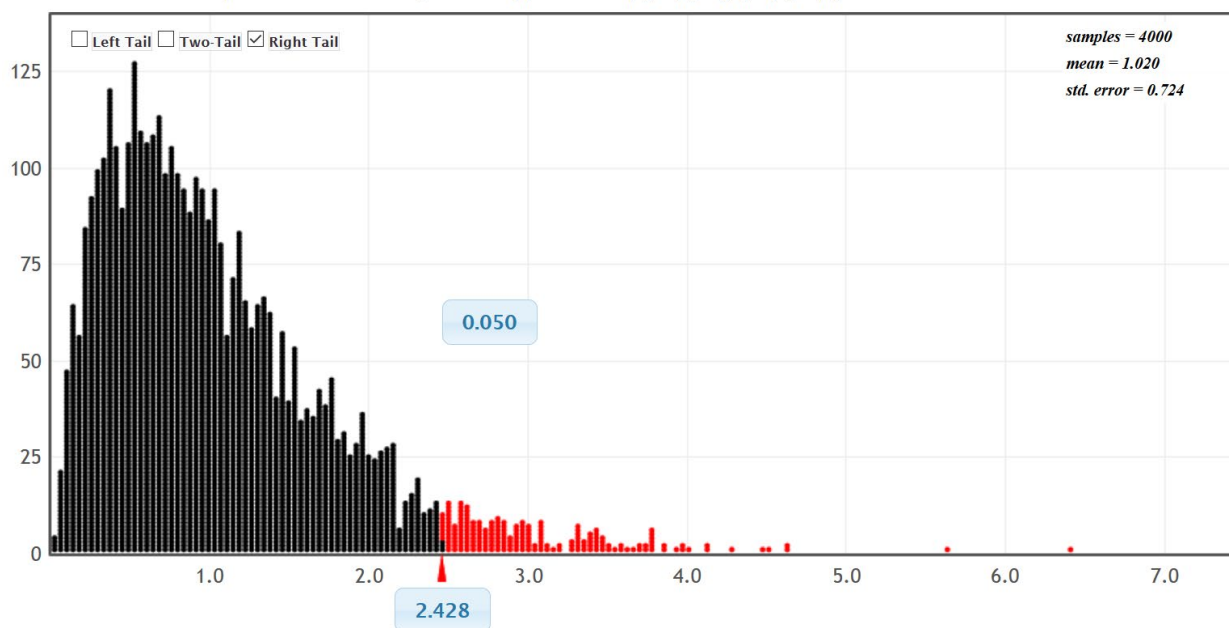
ANOVA Table

	df	SS	MS	F
Groups	4	10484530.0	2621132.5	7.922
Error	170	56249274.8	330878.1	
Total	174	66733804.8		

Notice the null hypothesis is that all five population means are equal. To simulate the null hypothesis click "Generate 1000 Samples" a few times. In simulations with only one or two groups, we usually use the sample mean, sample mean difference, sample proportion, or sample proportion difference. In tests with more than two groups, we cannot use that approach. When a test involves three or more groups, we will resort to using the test statistic to summarize the sample data.

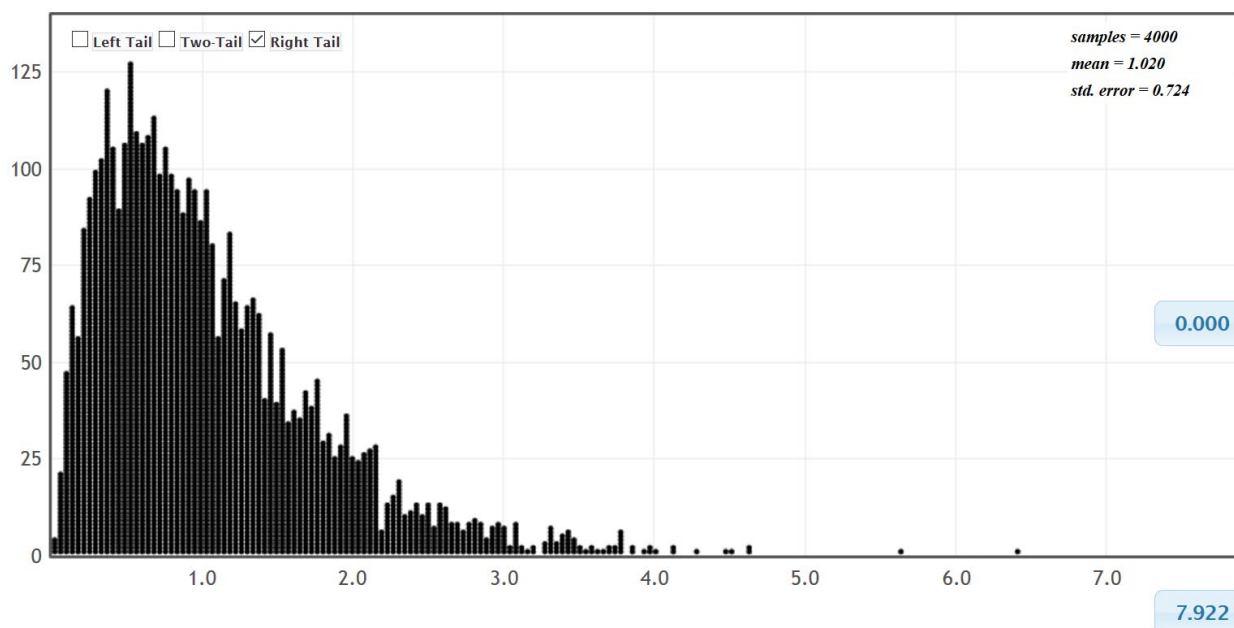
In this simulation, the computer has randomly collected thousands of samples and calculated thousands of F test statistics. Remember the real test statistic can be found under "Original Sample". ANOVA is a right tailed test so we will click on "Right-Tail". If we put in the 5% significance level in the proportion box in the right tail, we will have the critical value. Because of sampling variability, you will get slightly different answers, but this simulation gave a critical value of 2.428, which is not far from the theoretical critical value calculated by Statcato earlier. We can now use this graph to determine if the test statistic falls in the tail. Notice our F-test statistic of 7.922 does fall in the tail, so our sample data significantly disagrees with the null hypothesis and our variance between the groups is significantly higher than the variance within the groups.



Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ 

↑
F = 7.922

We can also calculate the P-value by putting the test statistic in the bottom box of the simulation. Notice our P-value came out to be about zero. So this sample data is unlikely to occur because of sampling variability if the null hypothesis was true. We would reject the null hypothesis and get the same conclusion as we did with the traditional approach.

Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ 

Problems Section 4B

(#1-10) Use each of the following ANOVA F-test statistics and the corresponding critical values to fill out the table.

	F-test stat	Sentence to explain F-test statistic.	Critical Value	Does the F-test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	+5.573		+2.886		
2.	+1.192		+3.113		
3.	+0.664		+2.949		
4.	+4.415		+3.125		
5.	+3.718		+4.117		
6.	+0.991		+2.009		
7.	+2.652		+1.875		
8.	+1.585		+3.225		
9.	+2.447		+2.798		
10.	+8.133		+2.891		

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.186			10%			
12.	0.0042			1%			
13.	2.59×10^{-4}			5%			
14.	0.006			1%			
15.	0.353			5%			
16.	0			10%			
17.	0.041			5%			
18.	0.274			10%			
19.	1.04×10^{-8}			1%			
20.	0.067			5%			

21. The F-test statistic compares the variance between the groups to the variance within the groups. Explain how the variance between the groups is calculated and what it tells us. Explain how the variance within the groups is calculated and what it tells us. How can we use the variance between and the variance within to calculate the F-test statistic?

22. If the variance between the groups were significantly larger than the variance within, would the F-test statistic be large or small? Explain why.

23. If the variance between the groups were about the same as the variance within, would the F-test statistic be large or small? Explain why.

24. The ANOVA printout involves the degrees of freedom within the groups, the degrees of freedom between the groups and the total degrees of freedom. How are the different degrees of freedom calculated?



(#25-28) Directions: Use the following Statcato statistics, graphs and ANOVA printout to test the population claims. For each of the following problems answer the following.

- Give the null and alternative hypothesis.
- Check the assumptions for a One-Way ANOVA test.
- Write a sentence to explain the F test statistic.
- Use the F test statistic and Critical Value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.
- Use the P -value and Significance Level to answer the following: Could the sample data or more extreme have occurred because of sampling variability or is it unlikely that the sample data occurred because of sampling variability? Explain your answer.
- Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- Write a conclusion for the hypothesis test addressing evidence and the claim.
- What is the variance between the groups? What is the variance within the groups? Was the variance between significantly higher than the variance within? Explain how you know.
- Was the categorical and quantitative variables related or not. Explain your answer.

25. A random sample of black bears were weighed at various times of the year. Some of the bears were weighed in the spring, some in the summer and some in the fall. The bears were tagged so that the same bear was not measured more than once. Use a 1% significance level and the following Statcato statistics, graphs and ANOVA printout to test the population claim that the time of year (season) is related to the weight of the bears.

One-way ANOVA: Significance level = 0.01

Selected column variables: C1 Spring Bear Weig... C2 Summer Bear Weig... C3 Fall Bear Weight...

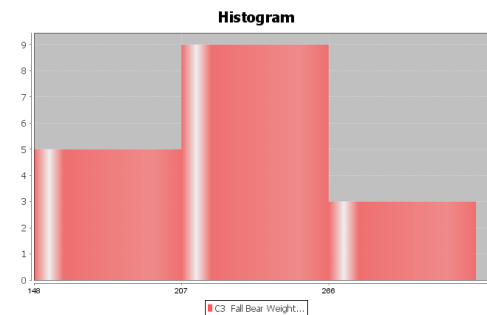
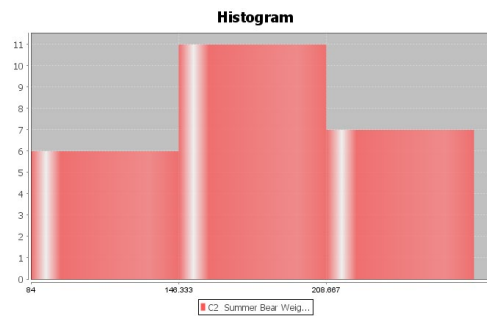
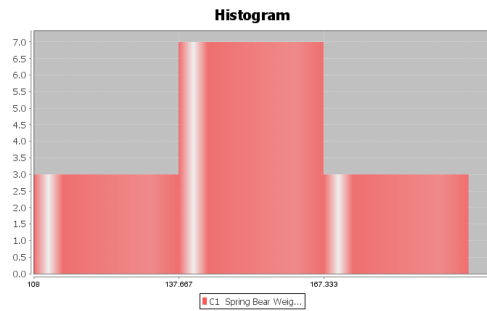
Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	2	45539.29263	22769.64632	13.55345	5.0472	0.00002
Error (Within Groups)	51	85679.46663	1679.98954			
Total	53	131218.75926				

Descriptive Statistics

Variable	Mean	Standard Deviation
C1 Spring Bear Weights in Pounds	151.385	22.463
C2 Summer Bear Weights in Pounds	182.125	48.017
C3 Fall Bear Weights in Pounds	228.118	40.769

Variable	N total
C1 Spring Bear Weights in Pounds	13
C2 Summer Bear Weights in Pounds	24
C3 Fall Bear Weights in Pounds	17





26. A census of Math 075 pre-stat students was taken in the fall 2015 semester. The students were separated into three sleep groups: low amount of sleep, moderate amount of sleep, high amount of sleep. They were also asked how many total units they have completed at the college. Though the data was not random, you can assume it was representative of Math 075 students at COC. Use a 10% significance level and the following Statcato statistics, graphs and ANOVA printout to test the claim that sleep is not related the total number of units completed.

One-way ANOVA: Significance level = 0.1

Selected column variables: C5 COC Units - Low ... C6 COC Units - Medi... C7 COC Units - High...

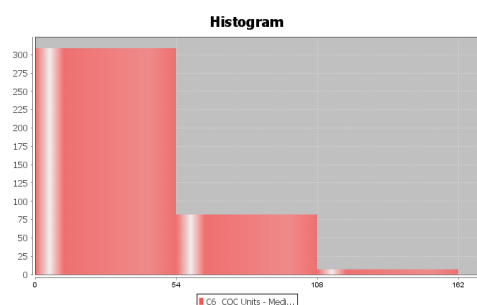
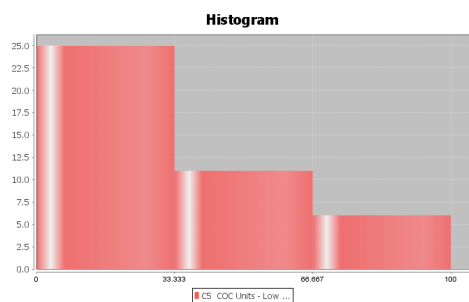
Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	2	2822.35625	1411.17813	1.83387	2.3133	0.16087
Error (Within Groups)	497	382446.38503	769.50983			
Total	499	385268.74128				

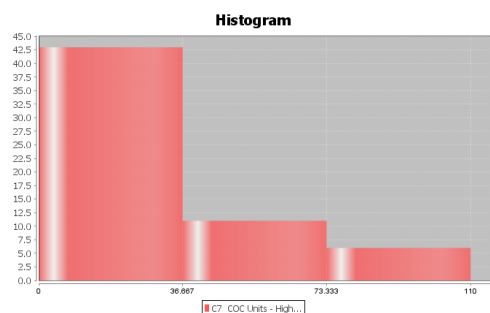


Descriptive Statistics

Variable	Mean	Standard Deviation
C5 COC Units - Low Sleep Group	32.952	28.586
C6 COC Units - Medium Sleep Group	32.990	27.585
C7 COC Units - High Sleep Group	25.675	28.178

Variable	N total
C5 COC Units - Low Sleep Group	42
C6 COC Units - Medium Sleep Group	398
C7 COC Units - High Sleep Group	60





27. A census of Math 075 pre-stat students was taken in the fall 2015 semester. The students were separated into four political parties: democratic, republican, independent party, and other political party. They were also asked number of alcoholic beverages they consume per week. Though the data was not random, you can assume it was representative of Math 075 students at COC. Use a 5% significance level and the following Statcato statistics, graphs and ANOVA printout to test the claim that political party is not related to the number of alcoholic beverages.

One-way ANOVA: Significance level = 0.05

Selected column variables: C9 # Drinks per Wee... C10 # Drinks per Wee... C11 # Drinks per Wee... C12 # Drinks per Wee...

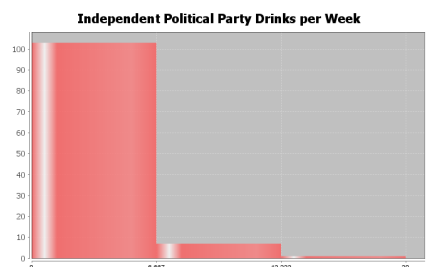
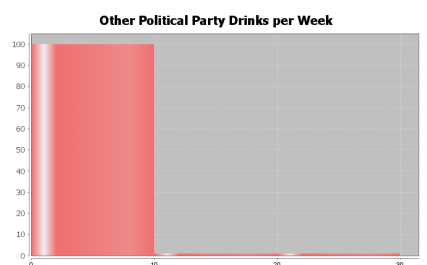
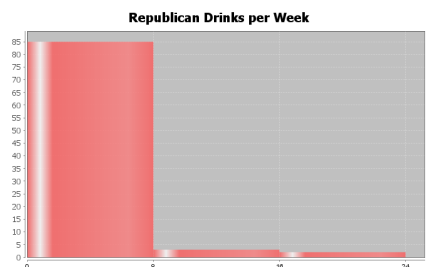
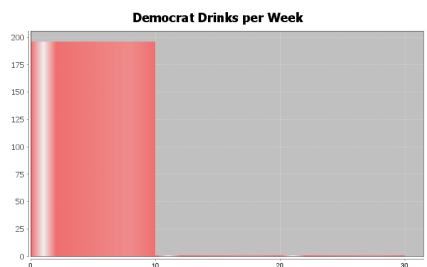
Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	3	25.44137	8.48046	0.89597	2.6228	0.44306
Error (Within Groups)	497	4704.16342	9.46512			
Total	500	4729.60479				

Descriptive Statistics

Variable	Mean	Standard Deviation
C9 # Drinks per Week - Democrats	0.914	2.566
C10 # Drinks per Week - Independent Political Party	1.342	2.943
C11 # Drinks per Week - Other Political Party	1.373	3.447
C12 # Drinks per Week - Republicans	1.411	3.753

Variable	N total
C9 # Drinks per Week - Democrats	198
C10 # Drinks per Week - Independent Political Party	111
C11 # Drinks per Week - Other Political Party	102
C12 # Drinks per Week - Republicans	90





28. A census of Math 075 pre-stat students was taken in the fall 2015 semester. The students were asked what their favorite social media is: Facebook, Instagram, Snapchat, or Twitter. They were also asked number minutes per day spent on social media. Though the data was not random, you can assume it was representative of Math 075 students at COC. Use a 5% significance level and the following Statcato statistics, graphs and ANOVA printout to test the claim that the type of social media is related to the number of minutes per day spent on social media.

One-way ANOVA: Significance level = 0.05

Selected column variables: C14 Facebook - Social Media, C15 Instagram - Social Media, C16 Snapchat - Social Media, C17 Twitter - Social Media

Source of Variation	DOF	SS	MS	Test statistic F	Critical value F	p-Value
Treatment (Between Groups)	3	169375.54058	56458.51353	8.20214	2.6354	0.00003
Error (Within Groups)	293	2016833.12272	6883.38950			
Total	296	2186208.66330				

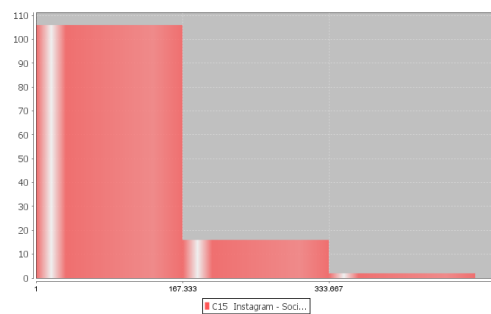
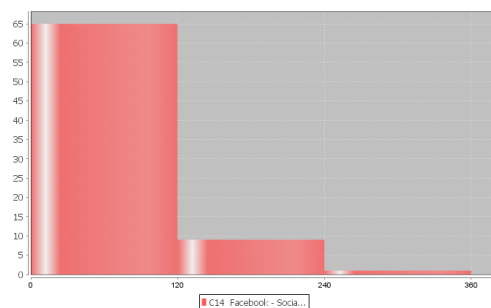


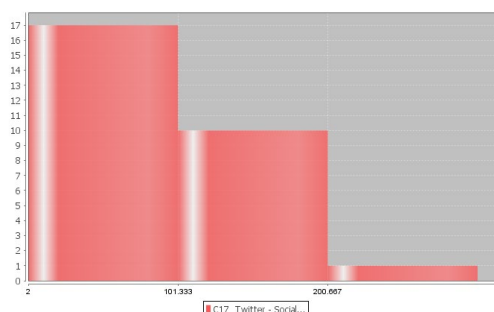
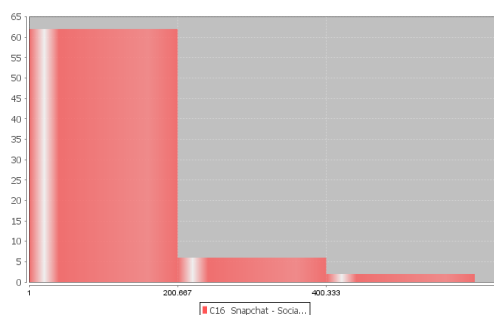
This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-By" [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

Descriptive Statistics

Variable	Mean	Standard Deviation
C14 Facebook - Social Media Minutes per day	43.867	58.103
C15 Instagram - Social Media Minutes per day	83.206	82.817
C16 Snapchat - Social Media Minutes per day	110.914	109.552
C17 Twitter - Social Media Minutes per day	90.964	59.408

Variable	N total
C14 Facebook - Social Media Minutes per day	75
C15 Instagram - Social Media Minutes per day	124
C16 Snapchat - Social Media Minutes per day	70
C17 Twitter - Social Media Minutes per day	28





(#29-33) Directions: Go to www.lock5stat.com and click on the StatKey button. Under the “More advanced randomization tests” menu click on “ANOVA for Difference in Means”. For each of the following problems, use a randomized simulation to answer the following. Assume the data met the assumptions for an ANOVA hypothesis test. For each problem, answer the following questions.

- Give the null and alternative hypothesis.
- The F-test statistic is given under “Original Sample”. Write a sentence to explain the F test statistic.
- Simulate the null hypothesis and put the significance level in the right tail to calculate the critical value. What was the critical value? (Answers will vary.)
- Use the F test statistic and Critical Value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.
- Put in the test statistic into the right tail to calculate the P-value. What was the P-value? (Answers will vary.)
- Use the P-value and Significance Level to answer the following: Could the sample data or more extreme have occurred because of sampling variability or is it unlikely that the sample data occurred because of sampling variability? Explain your answer.
- Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- Write a conclusion for the hypothesis test addressing evidence and the claim.
- What is the variance between the groups? What is the variance within the groups? Was the variance between significantly higher than the variance within? Explain how you know.
- Was the categorical and quantitative variables related or not. Explain your answer.

29. Use the random car data and a 1% significance level to test the claim that the country a car is from is related to its gas mileage. Go to www.matt-teachout.org and open the random car data. Copy and paste the country and the miles per gallon columns next to each other in a new excel spreadsheet. The country should be on the left and the miles per gallon should be on the right. Then copy both columns together. Go to www.lock5stat.com and click on the StatKey button. Under the “More advanced randomization tests” menu click on “ANOVA for Difference in Means”. Click on the “Edit Data” button and paste the country and mpg columns into StatKey. Click on “Generate 1000 Samples” a few times and then “Right-Tail”. Put in the original sample F-test statistic in the bottom box to estimate the P-value. Complete the questions above.



30. Under the “ANOVA for Difference in Means” menu in StatKey, click on the button at the top left of the page and click on “Sandwich Ants”. We are studying the number of ants that are drawn to different kinds of food. In this data, we are looking at the mean average number of ants that come to three different types of sandwiches left out to spoil. Use a 5% significance level to test the claim that the number of ants is not related to the type of sandwich.

31. Use the random car data and a 10% significance level to test the claim that the country a car is from is not related to its horsepower. Go to www.matt-teachout.org and open the random car data. Copy and paste the country and the horsepower columns next to each other in a new excel spreadsheet. The country should be on the left and the horsepower should be on the right. Then copy both columns together. Go to www.lock5stat.com and click on the StatKey button. Under the “More advanced randomization tests” menu click on “ANOVA for Difference in Means”. Click on the “Edit Data” button and paste the country and horsepower columns into StatKey. Click on “Generate 1000 Samples” a few times and then “Right-Tail”. Put in the original sample F-test statistic in the bottom box to estimate the P-value. Complete the questions above.

32. Under the “ANOVA for Difference in Means” menu in StatKey, click on the pulse rate and award data. This data looks at the average pulse rates of those people that have won Olympic, Academy and Nobel awards. Use a 1% significance level to test the claim that the population mean average pulse rate is related to the type of award the person won.

33. Under the “ANOVA for Difference in Means” menu in StatKey, click on the Homes for Sale (price by state) data. This data looks at the average selling price of homes in four different states. Use a 10% significance level to test the claim that the population mean average home price is related to the state the home is sold in.

Section 4C – Proportion Relationships: Two-population Proportion Test

Sometimes we wish to determine if a specific percentage from categorical data is related to various groups (populations). If we only have two populations, we can use a two-population proportion hypothesis test with a Z-score test statistic. If we have three or more populations, we will need to use a more advanced test statistic called the chi-squared test statistic. This is sometimes called a “Goodness of Fit” test. The key idea is to ask the question if the population percentage is the same in the various groups or is it significantly different.

Two-Population Proportion Test for Proportion Relationships

There are different ways of writing the null and alternative hypothesis. A population proportion can be described with the Greek letter pi (π) or with a “p”. Remember equal proportions goes with the null hypothesis of “not related” while any difference between the proportions indicates a relationship.

$H_0: p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)

$H_A: p_1 \neq p_2$ The population % is related to a categorical variable (% is related to the groups)

As we learned in the last chapter, the alternative hypothesis determines the type of test. If the alternative hypothesis is greater than ($>$) it is a right-tailed test. If the alternative hypothesis is less than ($<$) it is a left-tailed test. If the alternative hypothesis is not equal (\neq) it is a two-tailed test. While some prefer to use \geq or \leq for the null hypothesis, I prefer not to because of relationship implications.

Two-Tailed Null and Alternative Hypothesis

$H_0: p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)

$H_A: p_1 \neq p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

$H_0: \pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)

$H_A: \pi_1 \neq \pi_2$ The population % is related to a categorical variable (% is related to the groups)



Right-Tailed Null and Alternative Hypothesis

$H_0: p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
 $H_A: p_1 > p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

$H_0: \pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
 $H_A: \pi_1 > \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Left-Tailed Null and Alternative Hypothesis

$H_0: p_1 = p_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
 $H_A: p_1 < p_2$ The population % is related to a categorical variable (% is related to the groups)

OR

$H_0: \pi_1 = \pi_2$ The population % is NOT related to a categorical variable (% is not related to the groups)
 $H_A: \pi_1 < \pi_2$ The population % is related to a categorical variable (% is related to the groups)

Assumptions

It is very important to always check the assumptions for a hypothesis test in order to make sure that our sample data is as unbiased as possible. Remember that biased data may lead to a wrong conclusion (type 1 or type 2 error). Since we are now using this test to determine relationships, we may also need to prove cause and effect. If that is the case, we will need to use random assignment instead of a random sample.

Assumptions for a Two-population Proportion Test for Relationship

1. Random: The sample categorical data either should be a random sample (*if proving there is relationship*) or have used random assignment (*if proving cause and effect*).
2. Large sample size: The sample categorical data should have at least ten success ($x \geq 10$) and at least ten failures ($n - x \geq 10$). *For example, there should be at least 10 people with congestive heart failure (CHF) in the sample from the U.S. and at least 10 people without CHF in the sample from the U.S. There should also be at least 10 people with CHF in the sample from the Australia and at least 10 people without CHF in the sample from Australia.*
3. Data values within each sample and between the samples should be independent of each other. If the data was collected from one sample then the assumption is just that data values within the sample should be independent. If the data was collected from more than one sample, then the data values between the samples should also be checked for independence. *For example, we should not have people in our samples that are family members or the same people measured twice. The sample from the U.S. should not be connected to the sample from Australia. For example, the congestive heart failure (CHF) data should not come from a company that has hospitals in both countries. In an experiment, we should not control confounding variables by using the same group of people measured multiple times. This would fail the independent individuals' assumption. Random assignment is a better option for controlling confounding variables.*

Note: Some statisticians like to use the chi-squared test statistic even if there are only two populations of interest. If that is the case, use the assumptions for the goodness of fit test.

Z-test statistic for two-population proportion tests

The Z-test statistic measures the number of standard errors that the sample proportion from group 1 (\hat{p}_1) is above or below the sample proportion from group 2 (\hat{p}_2). It is "above" when the Z-test statistic is positive and "below" when the Z-test statistic is negative. It can also be thought of as the number of standard errors that the difference between the sample proportions ($\hat{p}_1 - \hat{p}_2$) is from zero or some other claimed difference. Z-scores usually are significant around two standard errors, but it is always good to refer to the critical value or P-value when judging significance.



The formula below seems daunting to calculate. Remember, no one in data science calculates this by hand with a calculator. Always use a computer program like R or Statcato. In the two-population proportion Z test, we often use pooling (\bar{p}). Pooling the proportions is combines the two data sets together before calculating the proportion. We need to assume that the population proportions are equal in the null hypothesis in order to pool. For this reason, pooling is usually used for a two-population proportion hypothesis test and is not used in a two-population proportion confidence interval.

$$p\text{-pooled } (\bar{p}) = \frac{(x_1 + x_2)}{(n_1 + n_2)}$$

$$Z\text{-test statistic for two population proportion (pooled)} = \frac{(\text{sample 1} - \text{sample 2})}{\text{standard error}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\left(\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}\right)}}$$

While this formula looks daunting, it is only counting how many standard errors the sample proportion for group 1 is above or below group 2. The most important thing is not calculating. It is interpreting and explaining the test statistic.

Z-test statistic for two-population sentence: The sample proportion for group 1 is # of standard errors (above or below) the sample proportion for group 2.

Look at the following two-population proportion printout.

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.96, 1.96	-0.412	0.6800

The Z-score test statistic is -0.412 . Since it is negative, we know that the sample proportion for group 1 is lower than the sample proportion for group 2.

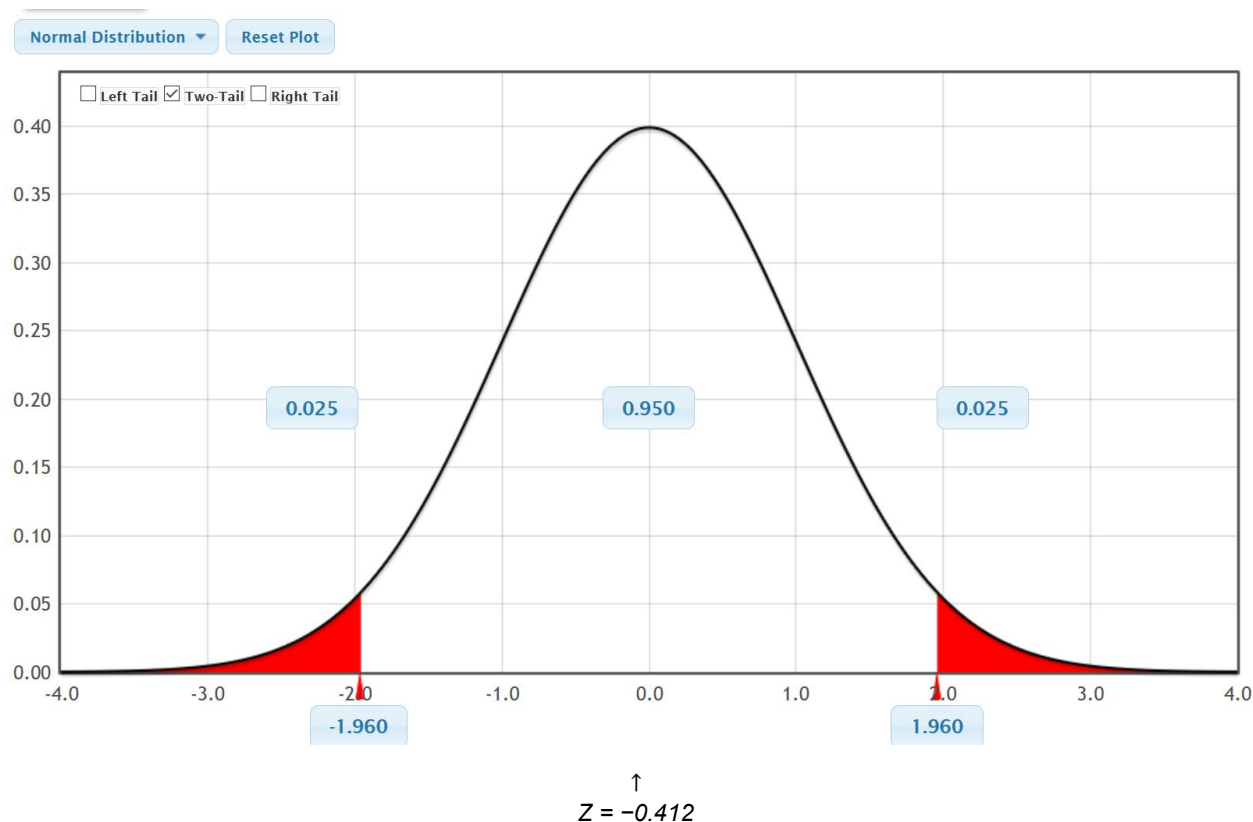
Z-test statistic sentence: The sample proportion for group 1 is 0.412 standard errors below the sample proportion for group 2. (Notice we did not say “ -0.412 standard errors”. A Z-score of -0.412 means “0.412 standard errors below”.)

Test statistics tell us if the sample data significantly disagrees with the null hypothesis. Remember the following rules.

- If the test statistic falls in one of the tails determined by the critical value or values, then the sample data significantly disagrees with the null hypothesis.
- If the test statistic falls does NOT fall in one of the tails determined by the critical value or values, then the sample data does NOT significantly disagree with the null hypothesis.

In the Statcato printout, the critical values are ± 1.96 . So the Z-test statistic does not fall in either tail. The sample data does not significantly disagree with the null hypothesis.





P-value

We also learned in the last chapter, that it is vital to know if the sample data occurred because of sampling variability (random chance). Remember, sample data always disagrees with the null hypothesis to some extent. The key is to determine why it is different. There are two reasons why the sample data disagrees with the null hypothesis. Maybe the null hypothesis is correct and the sample data disagrees because of sample data is always different (random chance). Another option is that the sample data disagrees because the null hypothesis is wrong. The key is that if you determine that it was not random chance, the only other option is that the null hypothesis is wrong. P-value is the key to making this decision about whether the data occurred by random chance. If the P-value is low (close to zero) then it is unlikely to be random chance. If the P-value is high, there is a possibility of the sample data occurring because of random chance.

- If $P\text{-value} \leq \text{significance level } (\alpha)$, then the sample data is unlikely to have occurred by random chance. Since sampling variability is ruled out, the null hypothesis must be wrong. So we “reject the null hypothesis”. A low P-value also indicates that the sample data significantly disagrees with the null hypothesis.
- If $P\text{-value} > \text{significance level } (\alpha)$, then the sample data is could have occurred by random chance. Since we do not know if sampling variability is involved or not, we also do not know if the null hypothesis is right or wrong. So we say we “fail to reject the null hypothesis” in this case. A high P-value also indicates that the sample data does not significantly disagree with the null hypothesis.

Randomized Simulation

In the last chapter, we saw that P-value could be calculated with randomized simulation or a randomization technique. This is a fabulous way for us to visualize what sampling variability (random chance) looks like if the null hypothesis is true. We have a computer create thousands of random samples under the premise that the null hypothesis is true. These simulated samples have the same sample sizes as the original sample data. If the real original sample data falls in the tail of the simulation it indicates that it is significant. The more in the tail the data is, the smaller the P-value. For a one-population proportion test, we see if the sample proportion is in the tail. For a two-population proportion test, we will see if the difference between the two sample proportions falls in the tail.



- If sample statistic falls in a tail of the simulation, then the sample data is significant and significantly disagrees with the null hypothesis.
- If the sample statistic does not fall in a tail of the simulation, then the sample data is not significant and does not significantly disagree with the null hypothesis.

Here is a chart from chapter four that summarizes test statistics, P-value and simulation.

	Significant Test Statistic	Test Statistic NOT Significant
	<i>(Test Statistic falls in tail determined by the critical value or values)</i>	<i>(Test Statistic does NOT fall in tail determined by the critical value or values)</i>
	OR	OR
	Small P-value	Large P-value
	<i>(P-value \leq significance level)</i>	<i>(P-value $>$ significance level)</i>
	OR	OR
	Sample Data in Tail	Sample Data NOT in Tail
	<i>(when simulating the Null Hypothesis)</i>	<i>(when simulating the Null Hypothesis)</i>
Is the sample data significantly different than H_0?	Yes. Significantly different	Not Significantly different
Could the sample data happen by random chance (sampling variability) if H_0 is true?	Unlikely	Could happen
Reject H_0 or Fail to Reject H_0?	Reject H_0	Fail to Reject H_0
Is there significant Evidence?	Yes. Is evidence	No evidence

Example (Two-Population Proportion Categorical Relationship Test)

Many high school and college students love to listen to music when they study. Some like to listen to their favorite music, while others just like the background noise. Use a 5% significance level to test the claim that liking the music is related to being able to memorize a large amount of information. A randomized experiment was done to test this claim. A group of college students were randomly assigned into two groups. Both groups had to memorize the same amount of information. The number of students that were able to memorize a significant amount of the information were classified as "high retention". One group listened to their favorite music and the other group had to listen to a type of music they hated. Confounding variables like the room environment and music volume were the same in both groups.

Label your variables.

p_1 : The percentage of college students that listen to liked music and can memorize a significant amount of information (high retention).

p_2 : The percentage of college students that listen to hated music and can memorize a significant amount of information (high retention).



Here is the sample data.

Liked Music: 25 total people, 10 high retention, $\hat{p}_1 \approx 0.4$

Hated Music: 24 total people, 11 high retention, $\hat{p}_2 \approx 0.458$

Sample Difference: $\hat{p}_1 - \hat{p}_2 \approx 0.4 - 0.458 = -0.058$

H_0 : $p_1 = p_2$ (The population % for high retention is NOT related to liking the music)

H_A : $p_1 \neq p_2$ (The population % for high retention is related to liking the music) CLAIM

(Notice this is a two-tailed test.)

Let us check the assumptions.

Is the sample data random or representative? Yes. The data was not a random sample of the population, but it was randomly assigned. So the sample data will not apply to all college students, but it has the capacity to prove cause and effect.

Is there at least 10 success and 10 failures in the sample data? Yes. In the liked music group, there were 10 high retention and 14 not high retention. In the hated music group, there were 11 high retention and 14 not high retention.

Are data values independent? It is difficult to know this without a detailed look at the people in the experiment. We did not have the same people measured twice, but instead used random assignment. We also should not have family members. If some of the students are friends and know each other, they may have similar taste in music. We will assume this data passes this assumption, but it might need further study.

Simulation Approach

Let us use StatKey to simulate the null hypothesis. Remember, the null hypothesis is equivalent to the difference being zero, so the simulation should be centered close to zero.

StatKey Directions for Two Population Proportions (percentages)

Randomization Hypothesis Tests → Test for Difference in Proportions → Under “edit data”, put in summary counts → click “generate 1000 samples” multiple times → click on tail determined by the alternative hypothesis → Enter sample proportion difference in bottom box. (If the difference is negative, put it in the left box. If the difference is positive, put it in the right box. P-value will be automatically calculated above the sample difference in the tail.)

We put the data into StatKey, simulated the null hypothesis, and then clicked on “two-tail”.

Edit data
✕

Please select values for two categories of count and sample size.

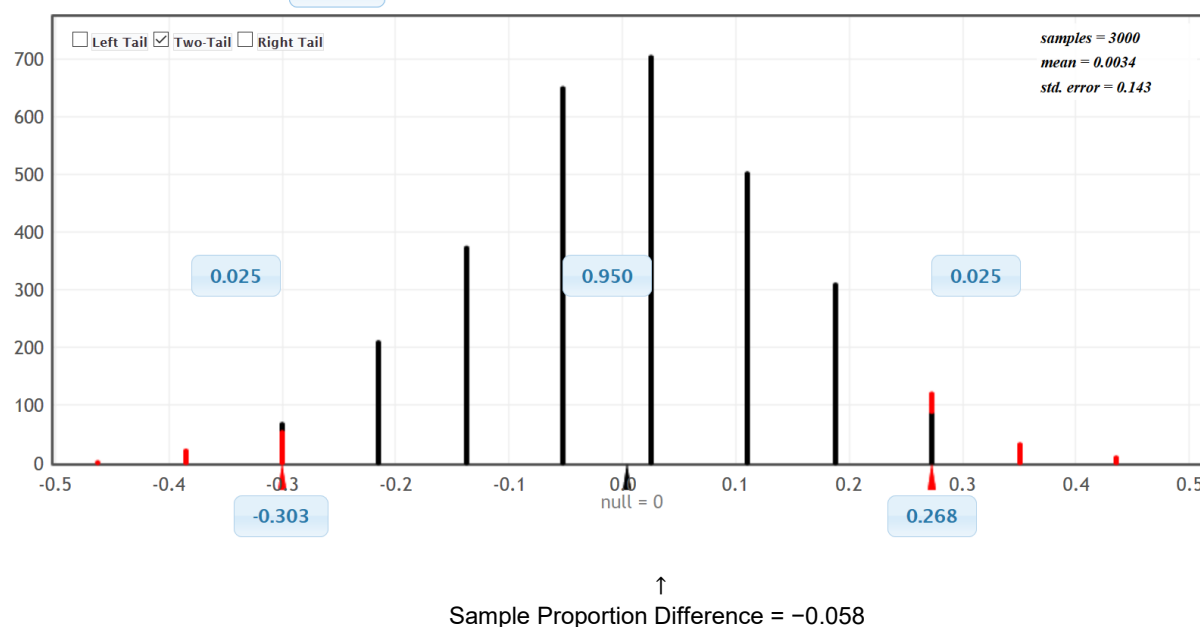
Group 1 count:	10
Group 1 sample size:	25
Group 2 count:	11
Group 2 sample size:	24

Ok



Original Sample

Group	Count	Sample Size	Proportion
Group 1	10	25	0.400
Group 2	11	24	0.458
Group 1-Group 2	-1	n/a	-0.058

Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ Null Hypothesis: $p_1 = p_2$ 

Notice that in simulation, it is important to identify the tail. With a 5% significance level and a two-tailed test, there is 2.5% in each tail. We see from the simulation that sample differences of approximately -0.303 or less are significant. Also sample differences of approximately +0.268 or higher are significant. Our real sample difference -0.058 was not in either of the tails. The sample data does not significantly disagree with the null hypothesis. It also tells us that the sample proportions are not significantly different.

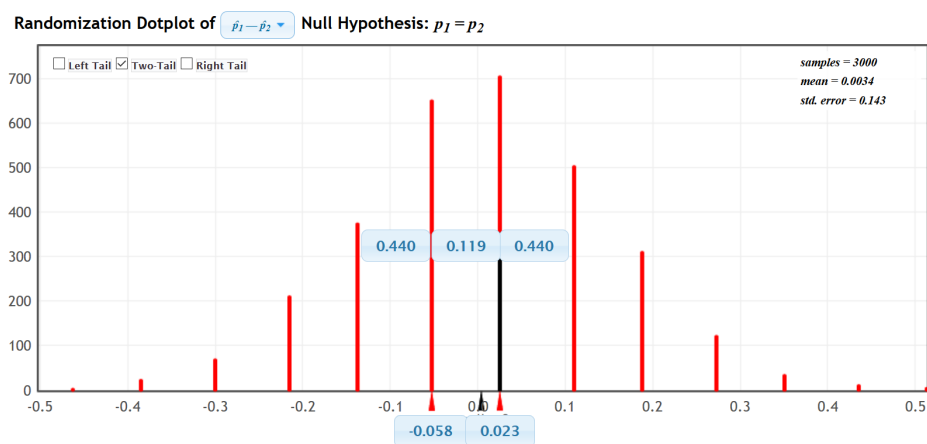
StatKey does not calculate the Z-score test statistic, but we do have the approximate standard error from the simulation of about 0.143. Using the test statistic formula, we get the following.

$$Z = \frac{(\text{sample proportion 1} - \text{sample proportion 2})}{\text{standard error}} = \frac{-0.058}{0.143} \approx -0.41$$

So the sample proportion of high retention for the liked music group was only 0.41 standard errors below the sample proportion of high retention for the hated music group. This is not significant. We do not have a critical value, yet we saw that the sample difference was not in the tail of the simulation.

Now let us use the simulation, to calculate the P-value and check whether this data could have happened because of sampling variability (random chance). If we enter the original sample difference of -0.058 in the left bottom box, we get the following.





Notice in a two-tailed test, you need to add the proportions in both tails (upper boxes) to get the P-value.

P-value $\approx 0.440 + 0.440 = 0.880 = 88.0\%$

P-value Sentence: If the null hypothesis is true, there is an 88.0% probability of getting this sample data or more extreme because of sampling variability.

Interpret the P-value: This is a very large P-value and is much larger than the 5% significance level. This indicates that the population proportions may be equal and the sample data could have happened because of sampling variability. Since sampling variability is involved, we must fail to reject the null hypothesis.

Conclusion: There is not significant evidence to support the claim that liking music is related to high retention. Notice that the alternative hypothesis (related) was the claim and we have a high P-value. Data seems to indicate they are not related, though we do not have significant evidence. This was an experiment with random assignment, so we may say the data indicates that liking the music does not cause a significant difference in the high retention percentage.

We could also use Statcato to calculate the test statistic, critical values and P-value.

Statcato Directions for Two Population Proportions (percentages)

Statistics \rightarrow Hypothesis Tests \rightarrow 2-Population Proportions \rightarrow Samples in one column, samples in two columns or summarized sample data \rightarrow put in alternative hypothesis sign (usually \neq for relationships)
 \rightarrow Hypothesized proportion difference: 0 \rightarrow check "use pooled estimate" \rightarrow put in significance level
 \rightarrow push "OK".

Here is the Statcato printouts for the same problem.



Hypothesis Test: 2-Population Proportions

Help

Inputs

☐ Samples in one column

Labels in column:

Values in column:

☐ Samples in two columns

Population 1:

Population 2:

☒ Summarized sample data

	Events	Trials
Population 1:	<input type="text" value="10"/>	<input type="text" value="25"/>
Population 2:	<input type="text" value="11"/>	<input type="text" value="24"/>

Significance

☒ Significance Level: 0 - 1.00 (e.g. 0.05)

☐ Confidence Level: 0 - 1.00 (e.g. 0.95)

Alternative Hypothesis

Alternative Hypothesis:

Hypothesized Proportion Difference:

☒ Use pooled estimate

OK Cancel

Hypothesis Test - Two population proportions: confidence level = 0.95

	Number of Events	Number of trials	Proportion
Sample 1	10	25	0.4
Sample 2	11	24	0.458

Null hypothesis: $p_1 - p_2 = 0.0$

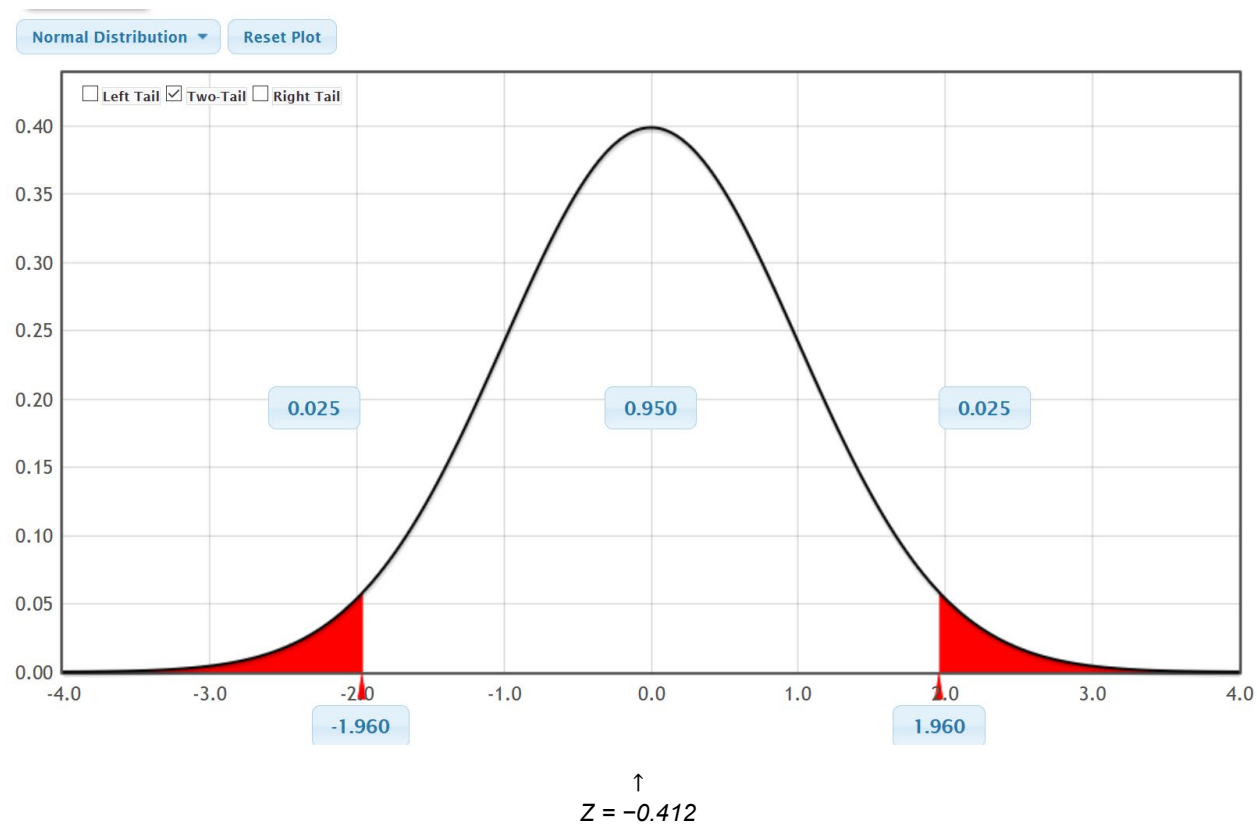
Alternative hypothesis: $p_1 - p_2 \neq 0.0$

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.96, 1.96	-0.412	0.6800

Notice the Z-test statistic is similar to what we got with StatKey, though the P-value is lower. Notice Statcato gave us critical values to compare the Z-test statistic to. The test statistic does not fall in the tail determined by the critical values. The P-value is still extremely large and indicates that if the null hypothesis was true, this data or more extreme could have happened because of sampling variability (random chance).

Also, notice the way Statcato wrote the null and alternative hypothesis. Saying two parameters are equal is the same as saying the difference is zero. You may see the null and alternative hypothesis written in different ways.





Problems Section 4C

(#1-10) Use each of the following two-population proportion Z-test statistics and the corresponding critical values to fill out the table.

	Z-test stat	Sentence to explain Z-test statistic.	Critical Value	Does the Z-test statistic fall in a tail determined by a critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	-1.835		± 1.645		
2.	+0.974		± 2.576		
3.	-1.226		-1.96		
4.	-3.177		± 1.96		
5.	+2.244		± 1.645		
6.	+1.448		± 2.576		
7.	-0.883		-2.576		
8.	+1.117		+1.96		
9.	+2.139		± 2.576		
10.	-0.199		-1.645		



(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P- value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.728			10%			
12.	0.0421			1%			
13.	2.11×10^{-4}			5%			
14.	0.0033			1%			
15.	0.176			5%			
16.	0			10%			
17.	0.0628			5%			
18.	0.277			10%			
19.	3.04×10^{-6}			1%			
20.	0			5%			

21. Explain the difference between random samples and random assignment.
22. List the assumptions that we need to check for a two-population proportion hypothesis test.
23. List the assumptions that we need to check for a two-population proportion hypothesis test that is using experimental design.
24. Explain how to use a two-population proportion hypothesis test to show that two categorical variables are related.
25. Explain how to use a two-population proportion hypothesis test to show there is a cause and effect between two categorical variables.

(#26-30) Directions: Use the following Statcato printouts to answer the following questions.

- a) Write the null and alternative hypothesis. Include relationship implications. Is this a left-tailed, right-tailed, or two-tailed test?
- b) Check all of the assumptions for a two-population proportion Z-test. Explain your answers. Does the problem meets all the assumptions?
- d) Write a sentence to explain the Z-test statistic in context.
- e) Use the test statistics and the critical value to determine if the sample data significantly disagrees with the null hypothesis. Explain your answer.
- f) Write a sentence to explain the P-value.
- g) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.
- h) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- i) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.
- j) Is the population proportion related to the categorical variable or not? Explain your answer.



26. The United States has the highest teen pregnancy rate in the industrialized world. In 2008, a random sample of 1014 teenage girls found that 326 of them were pregnant before the age of 20. In 2012, a random sample of 1025 teenage girls was taken and 334 were found to be pregnant before the age of 20. Let population proportion 1 represent 2008 and population proportion 2 represent 2012. Use a 10% significance level and the following Statcato printout to test the claim that the population percentage of teen pregnancies in the U.S. is lower in 2008 than it is in 2012. This claim would indicate that the population percentage of U.S. teen pregnancies is related to the year.

	Number of Events	Number of trials	Proportion
Sample 1	334	1025	0.326
Sample 2	326	1014	0.321

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	-1.645	-0.210	0.4168

27. While many Americans favor the legalization of marijuana, opponents of legalization argue that marijuana may be a gateway drug. They believe that if a person uses marijuana, then they are more likely to use other more dangerous illegal drugs. Use the table of random sample data given below and a 5% significance level to test the claim that marijuana users have a higher percentage of other drug use than non-marijuana users. This claim also would indicate that using Marijuana is related to using other drugs.

	Uses Other Drugs	Total
Uses Marijuana	87	213
Does not use Marijuana	26	219

	Number of Events	Number of trials	Proportion
Sample 1	87	213	0.408
Sample 2	26	219	0.119

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	1.645	6.850	$3.6839 \cdot 10^{-12}$



28. Use a 1% significance level and the following Statcato printout to test this claim that gender is not related to abstaining from drinking alcohol. If this is the case, then the percentage of men and women that do not drink alcohol should be the same. We took a random sample of 190 men and found that 66 of them never drink alcohol. We took a random sample of 250 women and found that that 137 of them never drink alcohol. We designated the proportion of men that never drink alcohol as population 1 and the women as population 2.

	Number of Events	Number of trials	Proportion
Sample 1	66	190	0.347
Sample 2	137	250	0.548

Significance Level	Critical Value	Test Statistic Z	p-Value
0.01	-2.576, 2.576	-4.182	$2.8937 \cdot 10^{-5}$



29. A health magazine claims that marriage status is one of the most telling factors for a person's happiness. Use a 10% significance level and the Statcato printout below to test the claim that the percent of married people that are unhappy is lower than the percent of single or divorced people that are unhappy. If this is the case, then perhaps being married, single or divorced is related to being unhappy. The following sample data was collected randomly. Population 1 represented married adults and population 2 represented single or divorced adults.

	Unhappy	Total
Married	74	200
Single or Divorced	97	200

	Number of Events	Number of trials	Proportion
Sample 1	74	200	0.37
Sample 2	97	200	0.485

Significance Level	Critical Value	Test Statistic Z	p-Value
0.10	-1.282	-2.325	0.0100

30. A tattoo magazine claimed that the percent of men that have at least one tattoo is greater than the percent of women with at least one tattoo. If this were true, then gender would be related to having a tattoo. Use a 5% significance level and the following Statcato printout to test this claim. A random sample of 857 men found that 146 of them had at least one tattoo. A random sample of 794 women found that 137 of them had at least one tattoo. Population 1 was the proportion of men with at least one tattoo and population 2 was the proportion of women with at least one tattoo.

	Number of Events	Number of trials	Proportion
Sample 1	146	857	0.170
Sample 2	137	794	0.173

Significance Level	Critical Value	Test Statistic Z	p-Value
0.05	1.645	-0.118	0.5468



(#31-34) Directions: go to www.lock5stat.com and click on StatKey. Then under the “Randomization Hypothesis Test” menu click on “Test for Difference in Proportions”. Create a randomized simulation of the null hypothesis to answer the following questions.

- a) Write the null and alternative hypothesis. Include relationship implications. Is this a left-tailed, right-tailed, or two-tailed test?
- b) What is the difference between the sample proportions? Adjust the tails of your simulation to reflect the significance level. Did your sample proportion difference fall in the tail?
- c) Does the sample data significantly disagree with the null hypothesis? Explain your answer.
- d) Put the sample proportion difference into the bottom box in the appropriate tail of your simulation in order to calculate the P-value. What was the P-value? (Answers will vary.) Write a sentence to explain the P-value.
- e) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.
- f) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- g) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.
- h) Is the population proportion related to the categorical variable or not? Explain your answer.
- i) Use the following formula to calculate the Z-test statistic. Write a sentence to explain the Z-test statistic in context. (Answers will vary.)

$$Z \text{ test stat} = \frac{\text{Sample Proportion Difference}}{\text{Standard Error}}$$

31. A body mass index of 20-25 indicates that a person is of normal weight for their height and body type. A random sample of 760 women found that 198 of the women had a normal BMI. A random sample of 745 men found that 273 of them had a normal BMI. A fitness magazine claims that the percent of women with a normal BMI is lower than the percent of men with a normal BMI. This would imply that gender is related to having a normal BMI. Let population 1 be the proportion of women with a normal BMI and population 2 be the proportion of men with a normal BMI. Use a 10% significance level and a randomized simulation in StatKey.

32. A new medicine has been developed that treats high cholesterol. An experiment was conducted and adults were randomly assigned into two groups. The groups had similar gender, ages, exercise patterns and diet. Of the 420 adults in the placebo group, 38 of them showed a decrease in cholesterol. Of the 410 adults in the treatment group, 49 of them showed a decrease in cholesterol. The FDA claims that the medicine is not effective in lowering cholesterol since the proportion for the placebo group and the treatment groups are about the same. Use a randomized simulation in StatKey, and a 1% significance level to test this claim.

33. A study was done to see if there is a relationship between smoking and being able to get pregnant. Two random samples of women trying to get pregnant were compared. A random sample of 135 women that smoke (population 1) found that 38 were able to get pregnant in the allotted amount of time. A random sample of 543 women that do not smoke (population 2) found that 206 were able to get pregnant in the allotted amount of time. Test the claim that the population percent of smoking women that were able to get pregnant is lower than the population percent of non-smoking women. This claim also implies that smoking is related to getting pregnant. Use a randomized simulation in StatKey, and a 5% significance level to test this claim.

34. A study was done to see if there is a relationship between the age of a person (teen or adult) and using text messages to communicate. A random sample of 800 teens (population 1) found that 696 of them use text messages regularly to communicate. A random sample of 2252 adults (population 2) found that 1621 of them use text messages regularly to communicate. Test the claim that population percentages are equal for the two groups implying that age is not related to using text messages. Use a randomized simulation in StatKey, and a 10% significance level to test this claim.



Section 4D – Proportion Relationships: Goodness of Fit Test

While the Z-score test statistic works well for two population proportion tests, it cannot handle proportions from three or more groups. For this case, we will introduce a new test statistic called “Chi-squared” (χ^2). This test statistic is usually used for more complicated categorical relationship analysis. The Goodness of Fit test works a lot like the two-population proportion relationship test except that there are now three or more groups. The opposite of three or more parameters being equal is not all of them being not equal. If even one is significantly different, we should reject the null hypothesis. For this reason, many statisticians prefer to use the phrase “at least one is not equal” or “the distribution is different than the null hypothesis”. I prefer the former.

Remember, if the population proportion or percentage is the same for all the groups, then it does not matter what group we are in. That would tell us that the population percentage is not related to the categorical variable that determines the groups. If the population proportion or percentage is different in at least one of the groups, then it does matter what group we are in. That would tell us that the population percentage is related to the categorical variable that determines the groups.

Null and Alternative Hypothesis for the Goodness of Fit Test

H_0 : $p_1 = p_2 = p_3 = p_4 = p_5$ The population % is NOT related to a categorical variable (% is not related to the groups)
 H_A : *At least one* \neq The population % is related to a categorical variable (% is related to the groups)

Expected Counts and Observed Counts

All hypothesis tests need to find some way of comparing the sample data to the null hypothesis. That is very difficult when you have three or more groups. The Goodness of Fit test compares the observed counts from the sample data to the expected counts from the null hypothesis. To calculate the Chi-Squared test statistic for a Goodness of Fit test, we will subtract the observed sample counts (number of successes) from each group to what we expect to happen if the null hypothesis was true (expected counts). Think of the observed counts as what really happened in the sample data and the expected counts as a theoretical count based on the null hypothesis being true. In this way, we can determine if the sample data significantly disagrees with the null hypothesis even if we have twenty groups.

Observed Counts: The counts from the sample data. Also called the number of successes or number of events.

Expected Counts: Theoretical counts based on the premise that the null hypothesis is true.

Calculating the Chi-squared test statistic (χ^2) for the Goodness of Fit Test

The Chi-squared test statistic works like a variance calculation. In fact, we have seen previously that the Chi-squared distribution is often used in one-population variance confidence intervals and hypothesis tests. Instead of calculating a sum of squares to measure the difference between data values and the mean, we will be calculating a sum of squares that measures the difference between the observed and expected counts. We need an average of the squares so we divide by the expected count for each group.

Chi-Squared Test Statistic formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$

The more groups you have in your data, the more difficult this formula is to calculate. While we will show an example of how the Chi-squared test statistic is calculated, it is always better to use a computer program to calculate it for you. It is more important to be able to explain the test statistic and be able to use it to determine if the sample data significantly disagrees with the null hypothesis.

Chi-Squared Test Statistic (χ^2) Sentence: The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis.

Degrees of Freedom for Goodness of Fit Test = $k - 1$



Interpreting the Chi-squared test statistic for a Goodness of Fit Test

The first thing to know about a Goodness of Fit test is that it is always a right-tailed test. It is never left-tailed or two-tailed. You may be comparing the proportions of twenty groups, but the Goodness of Fit test condenses it into one right-tailed test.

Degrees of Freedom

If we have ten groups, we will have ten expected counts and ten observed counts. So the degrees of freedom for our calculation will be the number of groups (k) minus one. For ten groups, the degrees of freedom will be $10 - 1 = 9$. This is important when looking up critical values.

Determining Significance

As with all hypothesis tests, if the test statistic falls in the tail determined by the critical value, then the sample data significantly disagrees with the null hypothesis. If the test statistic does not fall in the tail, then the sample data does not significantly disagree with the null hypothesis. The Goodness of Fit Test is a right-tailed test so the test statistic must fall in the right tail to be considered significant.

Assumptions for the Goodness of Fit Test for Categorical Relationships

1. Random: The sample categorical data should be either a random sample or representative (*if proving there is relationship*) or have used random assignment (*if proving cause and effect*).
2. Large sample size: The expected counts should be at least five. *In the Chi-squared test statistic calculation, we calculate theoretical counts based on the null hypothesis being true. These counts are called the expected counts (expected frequencies or expected values). In the Goodness of Fit test we want all of the expected counts to be five or greater. An expected count below five indicates the sample size was not large enough for a Goodness of Fit test.*
3. Data values within each sample and between the samples should be independent of each other. If the data was collected from one sample then the assumption is just that individuals should be independent. If the data was collected from more than one sample, then the samples and the individuals should be checked for independence. *As with the two population proportion assumptions, if we are doing an experiment, we should not control confounding variables by using the same group of people measured multiple times. This would fail the independent individuals' assumption. Random assignment is a better option for controlling confounding variables.*

Example 1 (Goodness of Fit Categorical Relationship Test) Case 1: Equal proportions but data collected from different samples with unequal sample sizes.

In the previous example, we looked at data comparing two groups, those that listened to a music they liked and those that listened to a music they hated. From this data, we were able to see if liking a music or not is related memorizing information (high retention).

The scientists in this experiment also had a third group that did not listen to any music. If you recall from our discussion about experimental design, this is called the control group.

Here is the sample data.

Liked Music: 25 total people, 10 high retention, $\hat{p}_1 \approx 0.4$

Hated Music: 24 total people, 11 high retention, $\hat{p}_2 \approx 0.458$

No Music: 26 total people, 19 high retention, $\hat{p}_3 \approx 0.731$

Let us use a 5% significance level to test the claim the having music or not is related to high retention.

H_0 : $p_1 = p_2 = p_3$ (High retention is NOT related to having music or not.)

H_A : At least one \neq (High retention is related to having music or not.) CLAIM

What are the expected counts?



To calculate the expected counts, we have to think about what we would expect to happen if the null hypothesis was true. Remember if there is no relationship between the variables, then the music should not matter when it comes to memorizing information. The percentage of high retention should be the same. So each of the three groups should have the same percentage and the same expected count. If we disregard music, then there was a total of 75 adults and 40 tested into high retention. So if the null hypothesis is true and music is not related to memorizing information, then all the music groups should have a proportion of $40/75 \approx 0.533$. In our study of categorical data analysis, we saw that to estimate an amount you multiple the proportion times the total number of people or objects in that group.

Expected Count for each Group = (proportion for group if null hypothesis was true) x (sample size of that group)

Expected Count for Liked Music Group = $0.533 \times 25 \approx 13.325$

Expected Count for Hated Music Group = $0.533 \times 24 \approx 12.792$

Expected Count for No Music Group = $0.533 \times 26 \approx 13.858$

Let us check the assumptions for this test. Notice that this is an experiment, so will require random assignment instead of random samples. Since the data was collected from multiple samples, we will need the groups and individuals to be independent.

1. Was the sample data collected randomly? **Yes.** The groups were randomly assigned in the experiment. This will account for confounding variables in the cause and effect study.

2. Are all of the expected counts at least five? **Yes.** The expected counts were 13.325, 12.792, and 13.858. All of them are greater than five.

3. Are the data values independent? **Yes.** It is always difficult to judge independence. Since the groups were randomly assigned and not the same people measured three times, the groups are probably independent of each other. It is difficult to judge whether the individuals are independent in an experiment. They are often volunteers and may have some relationship like maybe they all came from the same college. We will assume it passes this assumption for now, but may need to check with the people running the experiment.

Note: Z-test statistics can only compare two proportions at a time and cannot compare three or more proportions. Hence, we will need to use the chi-squared test statistic (χ^2).

Chi-Squared Test Statistic (χ^2)

The goal of any test statistic is to see if the sample data significantly disagrees with the null hypothesis. To do this, we compare the actual sample data “observed” counts to the theoretical “expected” counts if the null hypothesis was true.

We need to know if the observed counts for high retention (10, 11 and 19) are significantly different from the expected counts in the null hypothesis.

Observed Sample Count for Liked Music Group: 10 high retention

Observed Sample Count for Hated Music Group: 11 high retention

Observed Sample Count for No Music Group: 19 high retention

Expected Count for Liked Music Group = $0.533 \times 25 \approx 13.325$

Expected Count for Hated Music Group = $0.533 \times 24 \approx 12.792$

Expected Count for No Music Group = $0.533 \times 26 \approx 13.858$

Calculating Chi-Squared

We want to know if the expected counts were significantly different from the observed counts. This will tell us if the sample data significantly disagrees with the null hypothesis. The chi-squared test statistic will tell us.

Chi-Squared Test Statistic (χ^2): The sum of the averages of the squares of the differences between the observed sample counts and the expected counts if the null hypothesis was true.



When calculating the Chi-squared test statistic for the Goodness of Fit test, it is important that you pair the observed count with the correct expected count from that same group. For this reason, the observed and expected counts are labeled to reflect what group they came from.

Observed Sample Count for Liked Music Group: $O_1 = 10$

Observed Sample Count for Hated Music Group: $O_2 = 11$

Observed Sample Count for No Music Group: $O_3 = 19$

Expected Count from H_0 for Liked Music Group: $E_1 \approx 13.325$

Expected Count from H_0 for Hated Music Group: $E_2 \approx 12.792$

Expected Count from H_0 for No Music Group: $E_3 \approx 13.858$

Chi-Squared Test Statistic formula = $\sum \frac{(O-E)^2}{E}$

$$\chi^2 = \frac{(O-E)^2}{E} = \frac{(10-13.325)^2}{13.325} + \frac{(11-12.792)^2}{12.792} + \frac{(19-13.858)^2}{13.858} \approx 0.830 + 0.251 + 1.908 = 2.989$$

Interpreting Chi-Squared

Is the test statistic of $\chi^2 = 2.989$ significant? To judge if this is significant, we will need a critical value, P-value, or to see if it is in the tail of a simulation.

Note: Chi-squared Goodness of Fit test is always right tailed. Remember if the null hypothesis were true, then the observed and expected counts would be about the same. If that is the case then when you subtract them you should get about zero. So the center of the Chi-squared distribution should be close to zero. If you square numbers and add them up, it would be impossible for them to ever be negative.

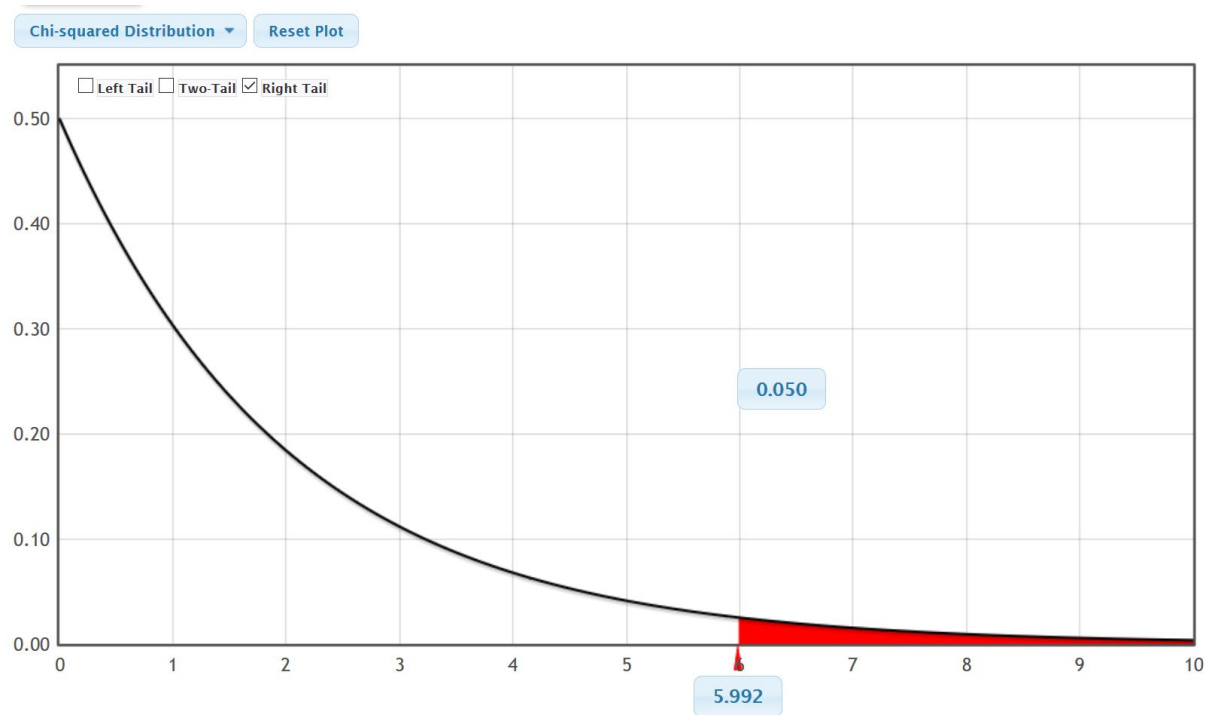
Degrees of Freedom: The chi-squared test statistic is based on the counts for the number of groups (k) so the degrees of freedom for a goodness of fit test is k-1. In this problem, there were three groups so the degrees of freedom is 3-1 = 2.

Since we have already calculated the test statistic, we can look up the critical value and P-value with the StatKey theoretical chi-squared function.

StatKey (Theoretical Chi-Squared)

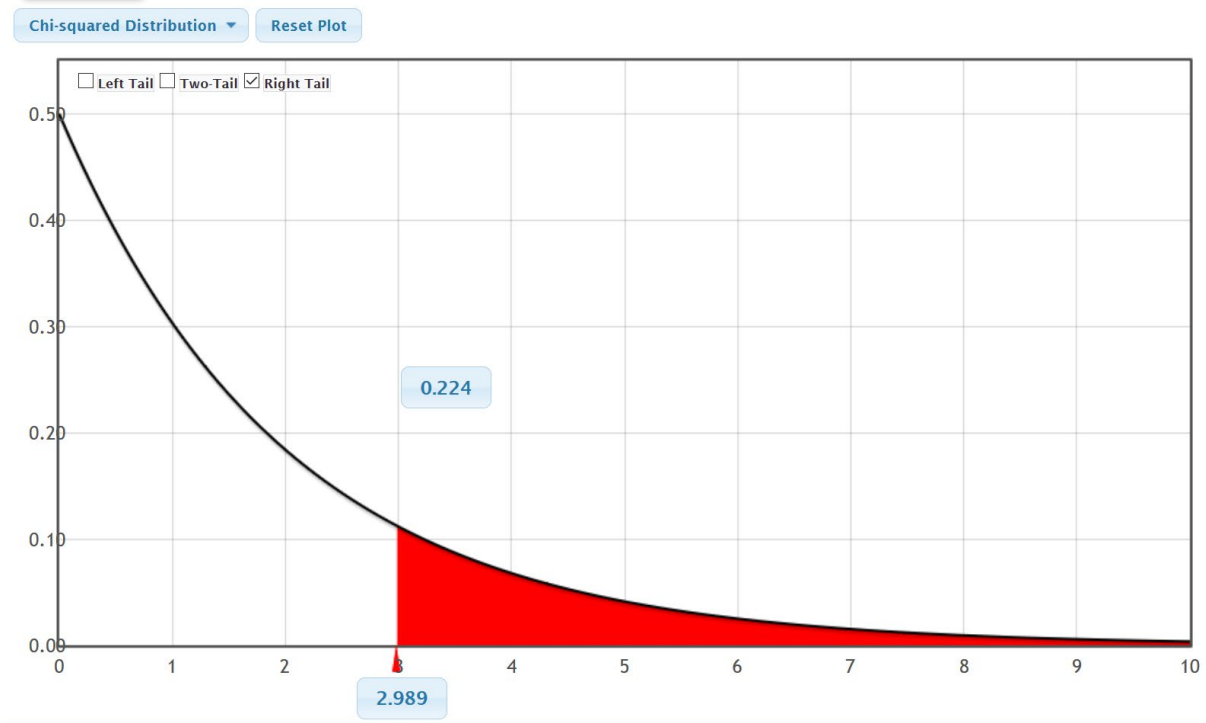
Theoretical Distributions → χ^2 → degrees of freedom: 2 → Click "right tail". (Remember chi-squared is always a right tailed test.) → To calculate the critical value, put in the significance level into the upper box. (The critical value will be below it.) → For the P-value, put the test statistic into the lower box. (The P-value will be above it.)





↑
 $\chi^2 = 2.989$

Notice that the critical value for a 5% significance level is 5.992. This means that the test statistic should be greater than 5.992 to be considered significant. Notice this implies that our chi-squared test statistic of 2.989 is not in the tail and not significant. So our sample data does not significantly disagree with the null hypothesis.



Notice that when we plugged in the test statistic of 2.989 into the theoretical Chi-squared curve, the estimated P-value is about $0.224 = 22.4\%$. This is a rather large P-value and is much larger than the 5% significance level. If the null hypothesis was correct, then this sample data or more extreme could have happened because of sampling variability (random chance).

Since we cannot rule out sampling variability, we should fail to reject the null hypothesis.

Conclusion: Since the P-value is high and the claim is the alternative hypothesis, our conclusion should be that we do not have significant evidence to support the claim that listening to music is related to high retention. The sample data indicates that listening to music is not related to high retention, though we do not have evidence.

We could also have calculated the test statistic, critical value and P-value with Statcato.

Statcato Directions for Goodness of Fit

First type in the observed counts in one column of Statcato and the expected counts into a second column. Take note of whether your expected counts are equal or not. In this case, they are not equal. The proportions were assumed equal in the null hypothesis, but since the sample sizes of the groups are different, the expected counts will be different.

	C1	C2
Var	Observed Counts Ex 1	Expected Counts Ex 1
1	10	13.325
2	11	12.792
3	19	13.858

Statistics → Multinomial Experiment → Chi-Square Goodness-of-Fit → Under (observed) Frequencies in Column: C1 (or whatever column has your observed sample counts) → Under Expected Frequencies: Click "Unequal Frequencies" → Under "Frequencies in column put in the column where you typed your expected counts → Put in the significance level → push OK.

Chi-Square Goodness of Fit Test

Help F1

Inputs

Observed Frequencies:

☒ Frequencies in Column: C1 Observed Counts Ex 1

Category names in Column: (optional)

☐ Categorical Data in Column:

Expected Frequencies:

☐ Equal Frequencies

☒ Unequal Frequencies

☒ Frequencies in Column: C2 Expected Counts Ex 1

☐ Probabilities in Column:

(assume in the same order as the categories provided)

☐ Categorical Data

Past Sample Data in Column:

Significance

Significance level: 0.05 0 - 1.00 (e.g. 0.05)

OK Cancel



Chi-Square Goodness-of-Fit Test:

Input: C1 Observed Counts Ex 1

Expected frequencies in C2 Expected Counts Ex 1

Category	Observed Frequency	Expected Frequency	Contribution to χ^2
0	10.0	13.325	0.8297
1	11.0	12.792	0.2510
2	19.0	13.858	1.9079

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
40.0	3	2	0.05	5.9915	2.9887	0.2244

Notice that our test statistic, critical value and P-value is about the same as the theoretical distribution in StatKey.

Example 2 (Goodness of Fit Categorical Relationship Test) *Case 2: Equal proportions from one sample.*

In the fall 2015 semester at COC, we asked the Math 140 statistics students what their favorite social media is. Here is the sample data. Use a 1% significance level to test the claim that the population proportions for each social media are not the same. This would indicate that the population percentage for social medias are related the type of social media. Notice the data came from one sample and has five types of social media. This means our sample size will be the same for each social media.

$H_0: p_1 = p_2 = p_3 = p_4 = p_5$ (The population proportion of COC statistics students that prefer a social media is not related to the type of social media.)

$H_A: \text{At least one } \neq$ (The population proportion of COC statistics students that prefer a social media is related to the type of social media.) CLAIM

AB
Which social media do you use the most?
Snapchat
Other
Facebook
Instagram
Facebook
Instagram
Other
Facebook
Facebook
Facebook
Snapchat
Twitter

	Count
Facebook	75
Instagram	124
Other	27
Snapchat	71
Twitter	31



Using Randomized Simulation

We can calculate the test statistic, critical value and P-value with a randomized simulation in StatKey. Like ANOVA, since there are three or more groups involved, we will not be able to put in the sample proportions directly into the simulation. Instead, the computer will use the Chi-squared test statistic to summarize the sample data.

Go to www.lock5stat.com and click on StatKey. Under the “More Advanced Randomization Tests” menu click on “ χ^2 Goodness of Fit”. Under “Edit Data”, type in the following. Do not forget to put a space after the comma. You can also use raw categorical data if you have it. Then push OK.

Choice, Count
 Facebook, 75
 Instagram, 124
 Other, 27
 Snapchat, 71
 Twitter, 31

Edit data
✕

Choice, Count
 Facebook, 75
 Instagram, 124
 Other, 27
 Snapchat, 71
 Twitter, 31

☐ Raw Data
☒ Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. For raw data, the file must have only one column. A summary counts table should contain two columns, where the first column contains categories and the second column contains counts.

Ok

OR



Edit data

Which social media do you use the most?

Snapchat
 Other
 Facebook
 Instagram
 Facebook
 Instagram
 Other
 Facebook
 Facebook
 Facebook
 Snapchat
 Twitter
 Snapchat
 Instagram
 Twitter
 Snapchat
 Instagram
 Facebook
 Snapchat
 Instagram

☒ Raw Data
☒ Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. For raw data, the file must have only one column. A summary counts table should contain two columns, where the first column contains categories and the second column contains counts.

Ok

It is good to look at the null hypothesis and make sure it is correct. Notice that saying that the proportions are equal is the same as saying each is 20% since we are dealing with one sample.

Edit Null Hypothesis

Edit the values below to update the null hypothesis.

*P*Facebook 0.2
*P*Instagram 0.2
*P*Other 0.2
*P*Snapchat 0.2
*P*Twitter 0.2

Ok (or hit Enter)

Notice that StatKey calculated the Chi-squared test statistic as $\chi^2 = 94.744$ under "Original Sample".



Original Sample

[Show Details](#)

$$n = 328, \chi^2 = 94.744$$

	Count
Facebook	75
Instagram	124
Other	27
Snapchat	71
Twitter	31

Let us check the assumptions for the Goodness of Fit Test. Under the “Original Sample” menu, click on “Show Details” to see the expected counts. Notice all of the expected counts are equal to 65.6. We can also see that the groups with the largest “Contribution to χ^2 ” have the most disagreement with the null hypothesis. Notice the largest “Contribution to χ^2 ” was 51.99, which came from the Instagram group.

Detailed Sample Table



	Count
Facebook	75 65.6 1.347
Instagram	124 65.6 51.99
Other	27 65.6 22.713
Snapchat	71 65.6 0.445
Twitter	31 65.6 18.249

Observed, Expected, Contribution to χ^2

Checking the Assumptions

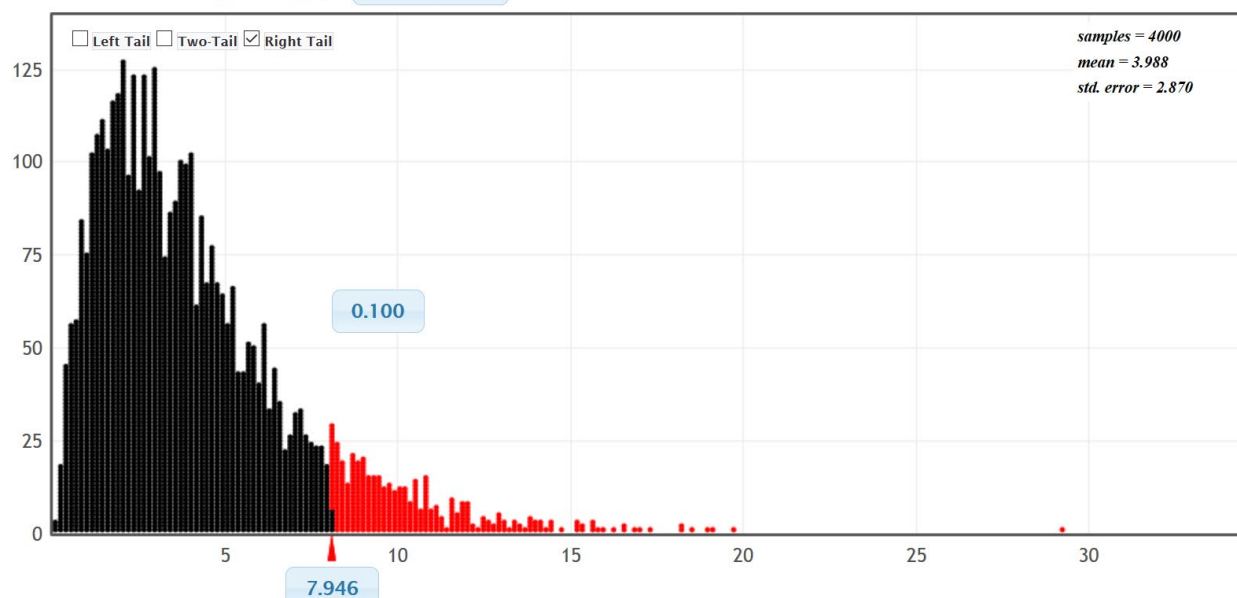
1. Is the sample data random or representative? **Yes.** Since the data was a census of all of the stat students in the fall 2015 semester, it is probably representative of all stat students at COC even though it is not a random sample.
2. Are the expected counts at least five? **Yes.** All of the expected counts were 65.6, which is greater than five.
3. Are the data values independent? **No.** Since this data came from one sample, we do not have to check that the samples are independent. It is difficult to judge if the individual stat students are independent or not. There are probably groups of friends or siblings in the data. In that case, they may have similar views about social media.

We can simulate the null hypothesis by clicking “Generate 1000 Samples” a few times. Notice the simulated distribution looks very skewed to the right. Remember the Goodness of Fit test is a right tailed test, so to calculate the Critical Value, click on “Right Tail” and put in the 10% significance level (0.10) in the tail proportion. The critical value came out to be approximately 7.946 in this simulation, but remember answers can vary due to sampling variability. In any case, our test statistic of 94.744 is way in the right tail.



Randomization Dotplot of χ^2 ,

Null Hypothesis

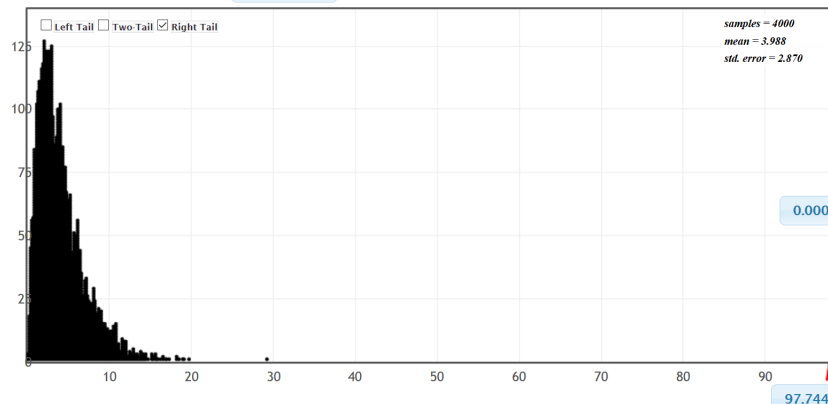


$$\chi^2 = 94.744$$

We can also calculate the P-value by plugging in the test statistic of 94.744. Remember not to confuse the simulated chi-squared values with the actual original sample test statistic. We have calculated 4000 Chi-squared values, but only the one under "Original Sample" is the real one based on the data. Our estimated P-value is zero.

Randomization Dotplot of χ^2 ,

Null Hypothesis



Original Sample

Show Details

 $n = 328, \chi^2 = 94.744$

	Count
Facebook	75
Instagram	124
Other	27
Snapchat	71
Twitter	31

Randomization Sample

Show Details

 $n = 328, \chi^2 = 16.482$

	Count
Facebook	75
Instagram	90
Other	51
Snapchat	56
Twitter	56

Since our test statistic fell in the tail of the simulation, we know the sample data significantly disagrees with the null hypothesis. Since our P-value was zero, we know it is highly unlikely that this sample data or more extreme occurred due to sampling variability if the null hypothesis was true.

Since our P-value was low, we will reject the null hypothesis.


Since our P-value was low and our claim was the alternative hypothesis, our conclusion should be that there is significant evidence to support the claim that the population proportions are related to the type of social media. However, remember that we may have failed one of the assumptions regarding independence.

We can also calculate the test statistic, critical value and P-value with Statcato.



In Statcato, we will go to the “Statistics” menu, and then click on “Multinomial Experiments” and “Goodness of Fit”. Since we are dealing with one sample and equal proportions, the expected counts will be equal. In that case, we can click on the equal (expected) frequencies button. We will still need to type in the observed counts or we can copy and paste the raw data. Put in our 10% significance level and push OK. Notice that the critical value, test statistic, and P-value are virtually the same as the simulation in StatKey.

Var	C1
Observed Counts Ex 2	
1	75
2	124
3	27
4	71
5	31

 Chi-Square Goodness of Fit Test
 ×

Help
F1

Inputs

Observed Frequencies:

☒ Frequencies in Column: C1 Observed Counts Ex 2

Category names in Column: (optional)

☐ Categorical Data in Column:

Expected Frequencies:

☒ Equal Frequencies

☐ Unequal Frequencies

☐ Frequencies in Column:

☐ Probabilities in Column:

(assume in the same order as the categories provided)

☐ Categorical Data

Past Sample Data in Column:

Significance

Significance level: 0.10 0 - 1.00 (e.g. 0.05)

OK
Cancel

Chi-Square Goodness-of-Fit Test:

Input: C1 Observed Counts Ex 2

Expected frequency = 65.6

Category	Observed Frequency	Expected Frequency	Contribution to χ^2
0	75.0	65.6	1.3470
1	124.0	65.6	51.9902
2	27.0	65.6	22.7128
3	71.0	65.6	0.4445
4	31.0	65.6	18.2494

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
328.0	5	4	0.10	7.7794	94.7439	0

Another Type of Goodness of Fit Test

To determine a proportion relationship with a Goodness of Fit test, the null hypothesis will be that the population proportions are equal. However, Goodness of Fit tests can also be used to determine if sample data fits a specific distribution of proportions that are not necessarily all equal. When the proportions are not equal in the null hypothesis, the expected counts will also not be equal.

Example 3: Goodness of Fit Test: Unequal proportions in the null hypothesis.

A famous example of using a Goodness of Fit test in this way occurred in the case of juries in Alameda County, USA. Juries are required to represent the racial demographic of their county, yet Alameda county juries were way out of compliance. Here is the racial demographic of Alameda county at the time of the scandal. This is our null hypothesis. We will use a 1% significance level and a Goodness of Fit test to test the claim that the juries were out of compliance with these proportion.

Edit Null Hypothesis✕

Edit the values below to update the null hypothesis.

<i>p_{White}</i>	<input style="width: 80%;" type="text" value="0.54"/>
<i>p_{Black}</i>	<input style="width: 80%;" type="text" value="0.18"/>
<i>p_{Hispanic}</i>	<input style="width: 80%;" type="text" value="0.12"/>
<i>p_{Asian}</i>	<input style="width: 80%;" type="text" value="0.15"/>
<i>p_{Other}</i>	<input style="width: 80%;" type="text" value="0.01"/>

Ok (or hit Enter)

$H_0 : p_1 = 0.54, p_2 = 0.18, p_3 = 0.12, p_4 = 0.15, p_5 = 0.01$ (Juries represent the Alameda racial demographic)
 $H_A : \text{at least one is } \neq \text{ (CLAIM) (Juries do NOT represent the Alameda racial demographic)}$



Detailed Sample Table

	Count
White	780 784.6 0.027
Black	117 261.5 79.88
Hispanic	114 174.4 20.895
Asian	384 217.9 126.509
Other	58 14.5 130.051

Observed, Expected, Contribution to χ^2

Original Sample

[Show Details](#)

$n = 1453$, $\chi^2 = 357.362$

	Count
White	780
Black	117
Hispanic	114
Asian	384
Other	58

Here is the sample data and Chi-squared test statistic. Notice the observed and expected counts are very different for African American, Hispanic American and Asian American.

Using a randomized simulation in StatKey, we see that the test statistic was in the tail and the P-value was zero.



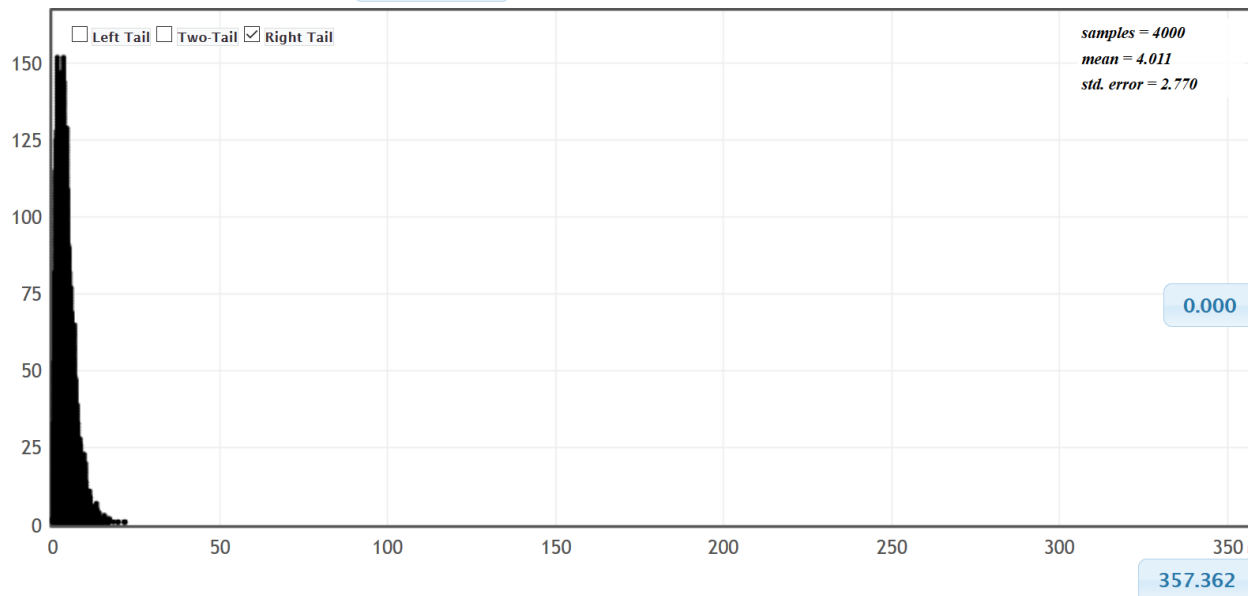
StatKey Chi-square Goodness-of-Fit

Alameda County Juries ▾ Show Data Table Edit Data Upload File Change Column(s)

Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Randomization Dotplot of χ^2 ,

Null Hypothesis



Hence we will reject the null hypothesis that the juries are in compliance and support the claim that the racial demographic of the juries in Alameda county were significantly out of compliance with the racial demographic of the county.

We can also calculate the test statistic and P-value with Statcato. Since the null hypothesis has specific proportions, we will need to type them in a column of Statcato. We will also need to type in the observed sample counts.

	C1	C2	
Var	Observed Counts Ex3	Ho Proportions Ex3	
1	780	0.54	
2	117	0.18	
3	114	0.12	
4	384	0.15	
5	58	0.01	

Now got to the “Statistics” menu in Statcato, click on “Multinomial Experiments” and then “Chi-Square Goodness of Fit”. We will need to enter the observed counts column under “Observed Frequencies”. Under “Expected Frequencies” click on “Unequal Frequencies” and then “Probabilities in Column”. Enter the column that has the proportions for the null hypothesis.



Chi-Square Goodness of Fit Test

Help F

Inputs

Observed Frequencies:

☒ Frequencies in Column: C1 Observed Counts Ex3

Category names in Column:

☐ Categorical Data in Column:

Expected Frequencies:

☐ Equal Frequencies

☒ Unequal Frequencies

☐ Frequencies in Column:

☒ Probabilities in Column: C2 Ho Proportions Ex3

(assume in the same order as the categories provided)

☐ Categorical Data

Past Sample Data in Column:

Significance

Significance level: 0.01 0 - 1.00 (e.g. 0.05)

OK Cancel

Chi-Square Goodness-of-Fit Test:

Input: C1 Observed Counts Ex3

Expected probabilities in C2 Ho Proportions Ex3

Category	Observed Frequency	Expected Frequency	Contribution to χ^2
0	780.0	784.62	0.0272
1	117.0	261.5400	79.8800
2	114.0	174.3600	20.8954
3	384.0	217.95	126.5088
4	58.0	14.5300	130.0510

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
1453.0	5	4	0.01	13.2767	357.3625	0

Notice the test statistic and P-value are the same as we calculated with StatKey.

Problems Section 4D

(#1-10) Use each of the following Goodness of Fit χ^2 -test statistics and the corresponding critical values to fill out the table.

	χ^2 -test stat	Sentence to explain χ^2 -test statistic.	Critical Value	Does the χ^2 -test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	+28.573		+9.117		
2.	+1.226		+7.113		
3.	+2.137		+5.521		
4.	+14.415		+6.114		
5.	+3.718		+7.182		
6.	+0.891		+3.994		
7.	+51.652		+14.881		
8.	+1.185		+4.181		



9.	+2.442		+8.619		
10.	+14.133		+10.336		

(#11-20) Use each of the following P -values and corresponding significance levels to fill out the table.

	P-value Proportion	P- value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.0006			10%			
12.	0.042			1%			
13.	9.16×10^{-7}			5%			
14.	0.739			1%			
15.	0.0035			5%			
16.	0			10%			
17.	0.419			5%			
18.	0.0274			10%			
19.	3.77×10^{-5}			1%			
20.	0.067			5%			

21. How is the degrees of freedom calculated in a Goodness of Fit test?

22. The χ^2 -test statistic compares the observed sample counts to the expected counts from H_0 . Explain how the expected counts are calculated.

23. Explain how the χ^2 -test statistic is calculated from the observed and expected counts.

24. If the observed sample counts were significantly different from the expected counts, would the χ^2 -test statistic be large or small? Explain why.

25. If the observed sample counts were close to the expected counts, would the χ^2 -test statistic be large or small? Explain why.

(#26-29) Directions: Use StatKey at www.lock5stat.com to simulate the following Chi-squared Goodness of Fit tests. Go to "more advanced randomization tests" at the bottom of the StatKey page. Click on the button that says " χ^2 Goodness of Fit". Under "Edit Data", type in the given sample data. Create a randomized simulation of the null hypothesis to answer the following questions.

a) Write the null and alternative hypothesis. Include relationship implications.

b) What is the degrees of freedom?

c) What is the Chi-squared test statistic? Write a sentence to explain the test statistic.

d) Adjust the right tail of your simulation to reflect the significance level. Did the Chi-squared test statistic fall in the tail?

e) Does the sample data significantly disagrees with the null hypothesis? Explain your answer.

f) Are the observed counts in the sample data significantly different from the expected counts from the null hypothesis? Explain your answer.

g) Put the Chi-squared test statistic into the bottom box in the right tail of your simulation in order to calculate the P -value. What was the P -value? (Answers will vary.) Write a sentence to explain the P -value.

h) Use the P -value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.

i) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.

j) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.



k) Is the population proportion related to the categorical variable or not? Explain your answer.

26. It is a big job to write and grade the AP-statistics exam for high school students each year. It is a difficult multiple-choice exam. All questions have five possible answers A-E. Use a 5% significance level and the following sample data to test the claim that percent of A answers is the same as the percent of B answers which is the same as C, D and E. This would indicate that the letter of the answer is not related to the percentage of times it happens. You can assume that the sample data meets the assumptions. Type the following sample data under the "Edit Data" menu of StatKey.

Choice, Count

A, 85

B, 90

C, 79

D, 78

E, 68

27. We collected data from all of the math 140 statistics students in the fall 2015 semester. A person that works at COC thinks that 80% of COC students drive alone, 10% carpool, 5% are dropped off by someone, 2% walk, 1% bike, and 2% use public transportation. Use a 5% significance level and the following sample data to test the claim that these percentages are wrong. You can assume that the data meets the assumptions for inference. Type in the proportions under "Null Hypothesis" in StatKey. Under "Edit Data" type in the following sample data from the fall 2015 survey data.

Choice, Count

Bicycle, 1

Carpool, 30

Drive Alone, 267

Dropped Off, 18

Public Transportation, 6

Walk, 10

28. We collected data from all of the math 140 statistics students in the fall 2015 semester. Use a randomized simulation in StatKey, a 5% significance level, and the following sample data to test the claim that the population percentages for the different political parties are different. This would indicate that the political party is related to the population percentages. You can assume that the data meets the assumptions for inference. Under "Edit Data", type in the following sample data from the fall 2015 survey data.

Choice, Count

Democratic, 110

Republican, 63

Independent, 65

Other, 90

29. Juries are required to meet the racial demographic of the county they represent. Here is the racial demographic for Alameda county: 54% Caucasian, 18% African American, 12% Hispanic American, 15% Asian American, and 1% other. We are worried that the juries in Alameda County may not be representing these percentages. Use randomized simulation, a 1% significance level, and the following observed sample counts to test the claim that the juries do not represent the demographic of the county. Under the "Edit Data" menu in StatKey, type in the following sample counts.



Jury Sample Data Observed Counts

Race, Count

Caucasian, 780

African American, 117

Hispanic American, 114

Asian American, 384

Other, 58

Under the "Null Hypothesis" menu, type in the following.

Edit Null Hypothesis✕

Edit the values below to update the null hypothesis.

<i>P</i> Caucasian	0.54
<i>P</i> African American	0.18
<i>P</i> Hispanic American	0.12
<i>P</i> Asian American	0.15
<i>P</i> Other	0.01

Ok (or hit Enter)

(#30-32) Directions: Use the following Statcato printouts to answer the following questions.

- a) Write the null and alternative hypothesis. Include relationship implications.
- b) Check the assumptions for a Goodness of Fit test.
- c) What is the Chi-squared test statistic? Write a sentence to explain the test statistic.
- d) Did the Chi-squared test statistic fall in the tail determined by the critical value?
- e) Does the sample data significantly disagrees with the null hypothesis? Explain your answer.
- f) Are the observed counts in the sample data significantly different from the expected counts from the null hypothesis? Explain your answer.
- g) What was the P-value? Write a sentence to explain the P-value.
- h) Use the P-value and significance level to determine if the sample data could have occurred by random chance (sampling variability) or is it unlikely to random chance? Explain your answer.
- i) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- j) Write a conclusion for the hypothesis test. Explain your conclusion in plain language.
- k) Is the population proportion related to the categorical variable or not? Explain your answer.



30. An online sports magazine wrote an article about the favorite sports in America. It said that 43% of Americans prefer Football, 23% of Americans prefer Baseball, 20% of Americans prefer Basketball, 8% of Americans prefer Hockey, and 6% of Americans prefer Soccer. When 130 randomly selected adults were asked their favorite sport, we found the following: 44 said Football, 26 said Baseball, 29 said Basketball, 13 said Hockey, and 18 said Soccer. Use a 5% significance level to test the claim that the proportions match the distribution claimed in the magazine article.

Chi-Square Goodness-of-Fit Test:

Input: C1 Observed Counts#4

Expected probabilities in C2 Null Hypothesis

Category	Observed Frequency	Expected Frequency	Contribution to χ^2
0	44.0	55.9	2.5333
1	26.0	29.9000	0.5087
2	29.0	26.0	0.3462
3	13.0	10.4	0.6500
4	18.0	7.8	13.3385

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
130.0	5	4	0.05	9.4878	17.3766	0.0016

31. Thousands of people die from car accidents across the U.S. every year, but is the day of the week related to the probability of having a fatal car accident? To test this claim, use a 1% significance level and a Goodness of Fit test to determine if the probabilities of a fatal car accident are significantly different. The following random sample data summary gives the observed number of the number of deaths from car accidents in the U.S. for each day of a randomly selected week. The total number of deaths for the week was 805.

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Number of Fatal Car Accidents	106	104	103	113	130	132	117

Chi-Square Goodness-of-Fit Test:

Input: C5 Observed #5

Expected frequency = 115.0

Category	Observed Frequency	Expected Frequency	Contribution to χ^2
0	106.0	115.0	0.7043
1	104.0	115.0	1.0522
2	103.0	115.0	1.2522
3	113.0	115.0	0.0348
4	130.0	115.0	1.9565
5	132.0	115.0	2.5130
6	117.0	115.0	0.0348

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
805.0	7	6	0.01	16.8119	7.5478	0.2731



32. The National Highway Traffic Safety Administration (NHTSA) publishes reports about motorcycle fatalities and helmet use. The following distribution shows the proportion of fatalities by location of injury for motorcycle accidents.

Location of Injury	Multiple Locations	Head	Neck	Thorax	Abdomen/Spine
Proportion	0.57	0.31	0.03	0.06	0.03

The random sample data below shows the distribution of 2068 randomly selected fatalities from riders that were not wearing a helmet. Use a 0.01 significance level to test the claim that the distribution for the sample does not match the proportions given by the NHTSA. Where is the largest discrepancy between the observed and expected value? What does this tell us about the importance of wearing helmets?

Location of Injury	Multiple Locations	Head	Neck	Thorax	Abdomen/Spine
Number of Deaths	1036	864	38	83	47

Chi-Square Goodness-of-Fit Test:

Input: C3 observed#6

Expected probabilities in C4 Ho#6

Category	Observed Frequency	Expected Frequency	Contribution to χ^2
0	1036.0	1178.76	17.2897
1	864.0	641.08	77.5150
2	38.0	62.04	9.3153
3	83.0	124.08	13.6006
4	47.0	62.04	3.6461

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
2068.0	5	4	0.01	13.2767	121.3667	0

Section 4E – Categorical Relationships: Contingency Tables

Vocabulary

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Contingency Table: Also called a two-way table. This table summarizes the counts when comparing two different categorical data sets each with two or more variables.

Marginal Percentage (Marginal Proportion): A single percentage or proportion without any conditions. In a contingency table, this can found with numbers in the margins.

Conditional Percentage (Conditional Proportion): The percentage or proportion calculated from a particular group or if a particular condition was true. These are the very important when studying categorical relationships.

Joint Percentage (Joint Proportion): A percentage or proportion involving two variables being true about the person or object, but does not have a condition. There are generally two types (AND, OR).



Introduction

An important field of exploration when analyzing data is the study of relationships between variables. A lot of thought has been put into determining which variables have relationships and the scope of that relationship. Is a person's diet related to having high blood pressure? Is the city a person lives in related to whether or not they have tuberculosis? Is being in a car accident related to texting while driving? These are all important questions that statisticians, data analysts and data scientists explore.

Relationships can be categorical \Leftrightarrow categorical, categorical \Leftrightarrow quantitative, and quantitative \Leftrightarrow quantitative. In this chapter, we will begin to explore the relationships between two categorical variables.

Remember, statistics is a deep well of mathematics and knowledge learned by years of study. There are much more advanced techniques for studying relationships, but we will be focusing on a basic introduction to the topic. You will find that a good understanding of this chapter will help tremendously when you go on to the more advanced techniques later on. For example, I find my students have many problems understanding the Chi-Squared distribution because they lack the foundational understanding of contingency (two-way) tables and analyzing differences between categories.

Note on Terminology: When studying relationships between variables you will hear different words used to describe the relationship. The most common are “relationship”, “association”, or “correlation”. “Correlation” is often used for describe a relationship between two quantitative variables (quantitative \Leftrightarrow quantitative), while “relationship” and “association” are used for two categorical variables (categorical \Leftrightarrow categorical) or for a categorical - quantitative relationship study (categorical \Leftrightarrow quantitative).

In this chapter, we will be using the terms “relationship” or “association”.

Note on Causation: One of the most famous statements in statistics is that “correlation is not causation”. Proving that one thing causes another is a much more complex kind of study and involves controlling confounding variables and experimental design. Remember that just because there is a relationship, that does not prove causation. There may be many other factors involved.

To analyze categorical data we need to know the counts (frequencies) for each categorical variable. This particularly important when you are studying categorical relationships. No data scientist or statistician finds the frequencies by hand. They use computer programs to make a contingency table (or two-way table).

Creating a contingency table with raw data and StatKey

Let us look at an example. Go to www.matt-teachout.org and click on the math 140 survey data fall 2015. We want to explore the relationship between the campus a person goes to and their political party.

First, we will need to check the data. When exploring relationships between two data sets, the data needs to be ordered pair. This usually means the data came from the same people. We also need to be careful of blanks. This means a person did not answer one or both of the questions. Start by copy and pasting the campus data and political party data into a fresh excel spreadsheet. A good rule of thumb is never mess up an original data set. Always copy and paste into a new excel file if you want to change things. The two columns of categorical data need to be in next to each other in the new Excel sheet. Otherwise, StatKey will not accept it. Go through the data and make sure there are no blanks. If there is a blank, delete that entire row. If you remember from chapter 1, this is called non-response bias. This process of deleting out missing cells is sometimes called “cleaning the data”.

To make a contingency table with StatKey, go to www.lock5stat.com and click the “StatKey” button. Now click on “Two Categorical Variables” under the “Descriptive Statistics and Graphs” menu. Then click on the “edit data” button. Copy both columns together in your excel spreadsheet and paste them into StatKey. Check the “raw data” box and the “data has header row” box and push “OK”.



Counts Table [Switch Variables](#)

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

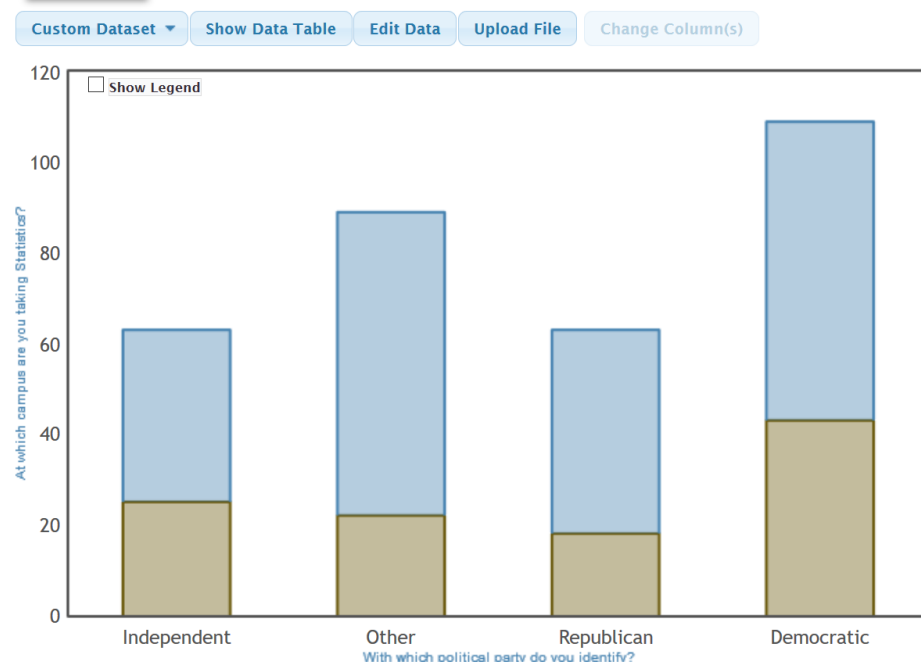
Proportions [Row](#) [Column](#) [Overall](#)

This is called a contingency table. Notice that we a lot of frequency information. Notice 66 is in the Democratic column and Valencia row, so 66 Math 140 students are both Democrat and go to the Valencia campus. Similarly, 18 math 140 students are both Republican and go to the Canyon Country campus.

The size of a contingency table is the number of rows by the number of columns. Totals are not included. This table has two rows (CCC and Valencia) and four columns (Independent, Other, Republican, and Democratic), so this is a “2 by 4” or “2×4” contingency table.

StatKey has several cool features with the contingency table. Notice it has created a stacked bar chart. This graph gives a visual representation of a contingency table. Notice if you place your cursor on any section of the graph the corresponding count lights up in the contingency table.

StatKey Descriptive Statistics for Two Categorical Variables



The “proportion” buttons are particularly useful. If we click on the “overall” proportion button. The computer calculates the intersection (AND) percentages for the entire data set. If we click on the “row” proportion button it gives conditional percentages for the rows. If we click on the “column” proportion button it gives the conditional percentages for the columns. We will discuss these more later, but these are very useful.



Proportions

Row Column Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Proportions

Row Column Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.231	0.204	0.167	0.398	1
Valencia Campus	0.176	0.31	0.208	0.306	1
Total	0.194	0.275	0.194	0.336	1

Proportions

Row Column Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.397	0.247	0.286	0.394	0.333
Valencia Campus	0.603	0.753	0.714	0.606	0.667
Total	1	1	1	1	1

Another feature is the “switch variables” button. Clicking on this button will switch the rows and columns.

Counts Table

Switch Variables

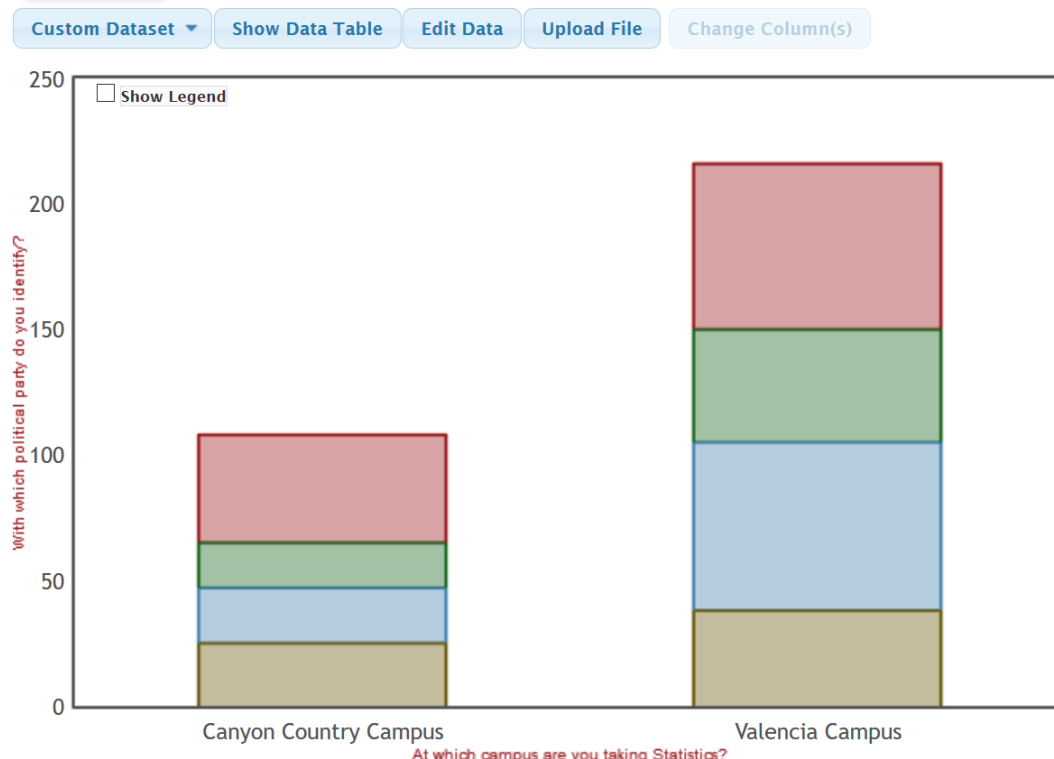
With which political party do you identify? \ At which campus are you taking Statistics?	Canyon Country Campus	Valencia Campus	Total
Independent	25	38	63
Other	22	67	89
Republican	18	45	63
Democratic	43	66	109
Total	108	216	324

Proportions

Row Column Overall



StatKey Descriptive Statistics for Two Categorical Variables



Notice that you can click on any section in the graph and it will highlight the count it came from in the contingency table. In addition, you can click on the proportion buttons to calculate and compare various proportions.

Creating a contingency table with summary counts and StatKey

Let us look at the same example again. As we said in the last section, categorical data is often not given in raw form. Sometimes a person may give you the summary counts (frequencies). In that case, you already have the contingency table, yet it is good to be able to put that into StatKey to create the stacked bar chart and use the switch variable and proportion features. To put in a contingency table into StatKey, go to www.lock5stat.com and click the "StatKey" button. Now click on "Two Categorical Variables" under the "Descriptive Statistics and Graphs" menu. Then click on the "edit data" button. Type in the table as seen below. Note that there should be a space after every comma and the totals are not included. There should also be a "[blank]" in the upper left corner. Uncheck the "raw data" box and check the "data has header row" box and push "OK". Notice this gives us the exact same table and graphs as if we had used the raw data.

[blank], Independent, Other, Republican, Democratic
 Canyon Country Campus, 25, 22, 18, 43
 Valencia Campus, 38, 67, 45, 66

Creating a contingency table with raw data and Statcato

You can also create a contingency table with Statcato. Copy and paste the ordered pair categorical data into a fresh excel spreadsheet. Make sure to clean the data and delete out any rows with missing values. Since this data set is over 300 values, go to the "edit" menu, "add multiple rows" and add another 100 rows. When that is done, copy and paste the two columns one at a time into Statcato. Statcato does not copy and paste multiple columns at the same time very well. It is best to copy and paste one at a time. Now go to the "Statistics" menu and click on "Multinomial Experiments". Now click on "Cross Tabulation and Chi-Square". Pick one column of data to be the row and the other



column of data as the column. Uncheck the box that says, "Perform chi-squared test". That is a more advanced analysis. Also, do not click on anything under the "frequency (optional)" menu. Now push "OK".

Statistics => Multinomial Experiments => Cross Tabulation => OK

If we use the campus and political party data from the previous example, we get the following from Statcato. Notice it gives the counts (frequencies), totals (All), and the intersection percentages (AND).

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Calculating Marginal Percentages

Marginal Percentages are percentages that involve only one variable and do not have a condition. They get their name because the amount and total are found in the margins (totals). Let us look at a couple examples. Remember a proportion and percentage can be found from the amount (frequency) and the total.

$$\text{Proportion} = \frac{\text{Amount (Frequency)}}{\text{Total}}$$

$$\text{Percentage} = \frac{\text{Amount (Frequency)}}{\text{Total}} \times 100\%$$

Example: Find the proportion and percentage of the math 140 students are democrat. Notice we need to find the amount of democrats and the total number of students. The amount of democrats will be in total part of the democrat row or column. The total number of students is often called the grand total and is found in the bottom right of the table.

$$\text{Proportion} = \frac{\text{Amount (Frequency)}}{\text{Total}} = \frac{109}{324} \approx 0.336$$

$$\text{Percentage} = \text{proportion} \times 100\% = 0.336 \times 100\% = 33.6\%$$

It is always better to use technology when you can instead of calculating something by hand. We could have found the proportion with StatKey by clicking on the "overall" proportion button. Statcato had this percentage already calculated as well. Notice the democrat data is summarized as a column. In both StatKey and Statcato, we need to look at the total in the democratic column to get the proportion. We can then convert the answer into a percentage or proportion as needed.

Proportions

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1



Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

Example: Use the tables above to give the proportion and percentage of the Math 140 students that attended the Canyon Country campus. Look for the Canyon Country campus data. Notice it is in the first row. So the number we are looking for is at the end of the first row under “total” or “All”.

Proportion of Math 140 students at the Canyon Country campus ≈ 0.333

Percentage of Math 140 students at the Canyon Country campus $\approx 33.3\%$

Calculating Joint Percentages

There are two types of joint percentages. The first type is the percentage of the total that has two things true about the person. We often call this the intersecting percentage or “AND”. The second type is the proportion or percentage of the total that has either one of two things true about the person. This is sometimes called the union percentage or “OR”. Intersecting percentages means that both things must be true about the person or object. Let us look at a few examples. Remember a proportion and percentage can be found from the amount (frequency) and the total.

Example: Find the proportion and percentage of the math 140 students that are both democrat AND attend the Valencia campus. Both things must be true about the person. In an “AND” (intersection) proportion, the amount can be found in the cell where the column and row meet. We will still use the “grand total” in the lower right corner as the total, since we need to include everyone in the data set. Look at the where the democratic column meets the Valencia row. There are 66 students that have both characteristics. This is the amount we need. The grand total is still 324 so here is the proportion and percentage calculation. Round your answer to three significant figures.

$$\text{“AND” Proportion} = \frac{\text{Frequency in intersection cell}}{\text{Grand Total}} = \frac{66}{324} \approx 0.2037037 \approx 0.204$$

$$\text{“AND” Percentage} = \text{proportion} \times 100\% = 0.204 \times 100\% = 20.4\%$$

Again, we could have used technology to get that answer. We could have found the proportion with StatKey by clicking on the “overall” proportion button. Statcato had this percentage already calculated as well. Both times, we need to look at the cell where the Democratic column meets the Valencia row.

Proportions

Row Column Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)



Example: Use the tables above to give the proportion and percentage of the Math 140 students that both attend the Canyon Country campus AND are Republican. Look for where the Canyon Country campus row meets the Republican Column.

Proportion of Math 140 students at the Canyon Country campus AND Republican ≈ 0.056

Percentage of Math 140 students at the Canyon Country campus AND Republican $\approx 5.6\%$

Example: Now calculate the proportion and percentage of Math 140 students that either are at the Valencia campus OR are Democratic. This means we need to include anyone that was Democrat regardless of campus and include anyone at the Valencia campus regardless of the political affiliation. This is a more difficult calculation. Here is a couple common formulas for “OR” (union) percentages.

$$\text{“OR” (Union) Proportion} = \frac{(\text{Row Total} + \text{Column Total} - \text{Intersection Cell})}{\text{Grand Total}} = \frac{(216 + 109 - 66)}{324} = \frac{(259)}{324} \approx 0.79938 \approx 0.799$$

It is better to use technology if we can. StatKey and Statcato printouts can help us calculate the “OR” (union) proportion or percentage.

“OR” (Union) Proportion = Row Total Proportion + Column Total Proportion – Intersection Cell Proportion

Notice these proportions are given in the StatKey table we can use them to calculate the “OR” proportion.

Proportions

Row Column Overall

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.077	0.068	0.056	0.133	0.333
Valencia Campus	0.117	0.207	0.139	0.204	0.667
Total	0.194	0.275	0.194	0.336	1

“OR” (Union) Proportion = Row Total Proportion (Valencia) + Column Total Proportion (Democratic) – Intersection Cell Proportion (where Valencia and Democratic meet)

$$= 0.667 + 0.336 - 0.204 = 0.799$$

(We can convert this proportion to a percentage if needed. Percent = Proportion $\times 100\% \approx 0.799 \times 100\% \approx 77.9\%$)

“OR” (Union) Percentage = Row Total % + Column Total % – Intersection Cell %

Notice these percentages are given in the Statcato table we can use them to calculate the “OR” percentage.

Cross Tabulation and Chi-Square

rows in C1 At which campus ..., columns in C2 With which polit...

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	43.0 (13.27%)	25.0 (7.72%)	22.0 (6.79%)	18.0 (5.56%)	108.0 (33.33%)
Valencia Campus	66.0 (20.37%)	38.0 (11.73%)	67.0 (20.68%)	45.0 (13.89%)	216.0 (66.67%)
All	109.0 (33.64%)	63.0 (19.44%)	89.0 (27.47%)	63.0 (19.44%)	324.0 (100.00%)

“OR” (Union) Percentage = Row Total % (Valencia) + Column Total % (Democratic) – Intersection Cell % (where Valencia and Democratic meet) = 66.7% + 33.6% – 20.4% = 79.9%



Conditional Proportions and Percentages

Conditional proportions and percentages are the key to understanding categorical relationships. A condition is thought of as prior knowledge about the person or situation that may change the percentage. Let us say that the Los Angeles Lakers have a 75% chance of beating the Phoenix Suns. If the Lakers best player LeBron James does not play, will that change the percentage? Of course. Knowing that LeBron James will not play is called a condition.

In contingency tables, a condition involves restricting to one particular group before you calculate the percentage.

Example: What percentage of the Canyon Country campus Math 140 students are Democrat?

First notice that this is not a joint proportion. It does NOT ask for the percentage of all students that are both Democrat and go to the Canyon Country campus.

The key is to identify which group we are restricting ourselves to. In other words, what is the condition? Look for words that say “if” or “given this is true” or “out of”. This designates the condition. In this example, notice that the problem said “of the Canyon Country students”. That means that we are supposed to only look at the Canyon Country students when we find our amount (frequency) and total. A commonly used method for calculating conditional percentages from a contingency table is to circle the row or column that has your condition (Canyon Country). Then only use numbers in that row or column.

Counts Table

[Switch Variables](#)

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions

[Row](#)
[Column](#)
[Overall](#)

Notice that the Canyon Country Campus counts are in the first row. So we should only use numbers in the first row. We should not use the grand total anymore. We need the total number of students that attend the Canyon Country campus. In other words, the total from our condition. The amount will be the number of democrats in the Canyon Country row. In other words the intersection cell frequency.

$$\text{Conditional Proportion} = \frac{\text{Amount in Intersection Cell (Canyon Country meets Democratic)}}{\text{Row or Column Total (Row total Canyon Country)}} = \frac{43}{108} \approx 0.398148 \approx 0.398$$

We can use the “row” and “column” proportion buttons in StatKey to find this conditional proportion. Since the condition is a row, we should click the “row” proportion button.

Proportions

[Row](#)
[Column](#)
[Overall](#)

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.231	0.204	0.167	0.398	1
Valencia Campus	0.176	0.31	0.208	0.306	1
Total	0.194	0.275	0.194	0.336	1

Notice the answer we are looking for is given in the intersecting cell. If we restrict ourselves to considering only the Canyon Country students, 0.398 or 39.8% of them are democrat.



Example: What proportion of the republican math 140 students attend the Valencia campus? To answer this we need to recognize that we are no longer considering all the students. We are restricting our proportion to considering only the republican students (“out of”). Since the condition is being republican, we should only use numbers in the republican column. The total will now be the total number of republicans and the amount will be the amount of republicans that attend the Valencia campus.

Counts Table

[Switch Variables](#)

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	25	22	18	43	108
Valencia Campus	38	67	45	66	216
Total	63	89	63	109	324

Proportions

[Row](#)
[Column](#)
[Overall](#)

$$\text{Conditional Proportion} = \frac{\text{Amount in Intersection Cell (Republican meets Valencia)}}{\text{Row or Column Total (column total Republican)}} = \frac{45}{63} \approx 0.7142857 \approx 0.714$$

We can also use StatKey to find what proportion of Republican Math 140 students attend the Valencia campus. Notice our condition is now republican (“out of”). This is a column so I will click the “column” proportion button in StatKey.

Proportions

[Row](#)
[Column](#)
[Overall](#)

At which campus are you taking Statistics? \ With which political party do you identify?	Independent	Other	Republican	Democratic	Total
Canyon Country Campus	0.397	0.247	0.286	0.394	0.333
Valencia Campus	0.603	0.753	0.714	0.606	0.667
Total	1	1	1	1	1

Notice now we want to restrict ourselves to the Republican column. The conditional proportion we are looking for is 0.714 or 71.4%.

Relationship Principle

Let us go back to the LeBron James example. The key to understanding categorical relationships is to judge how close or far apart conditional percentages are.

Chances of Lakers winning if LeBron James plays $\approx 75\%$

Chances of Lakers winning if LeBron James does not play $\approx 40\%$

These percentages are significantly different, so it tells us that the condition of LeBron James playing in the game is related to the Lakers winning.

Let us look at another example using the Lakers chances of beating the Phoenix Suns.

Chances of Lakers winning if it snows in Nebraska $\approx 75\%$

Chances of Lakers winning if it does not snow in Nebraska $\approx 75\%$

These percentages are not significantly different, so it tells us that the condition of snowing in Nebraska is not related to the Lakers winning. The condition does not matter.



Relationship Principle:

Close Conditional Percentages = Condition is NOT related to the categorical variable

Significantly Different Conditional Percentages = Condition IS related to the categorical variable

Note: You cannot compare any conditional percentages you want. They must be the same variable for the percentage and from different groups (different condition). You cannot compare the percentage of republicans from the Canyon Country campus to the percentage of democrats from the Valencia campus. They are not the same thing and will likely have very different percentages regardless of the relationship. Compare the percentage of republicans from the Canyon Country campus to the percentage of republicans from the Valencia campus. That will give us information about the relationship. Conditional percentage analysis is the basis behind the Chi-Squared test statistic we will learn in chapter 5.

Practice Problems Section 4E

1. If the proportions for a categorical variable from one group are significantly different from another group, what does that indicate about the relationship between that variable and the groups?
2. If the proportions for a categorical variable from one group are almost the same as another group, what does that indicate about the relationship between that variable and the groups?

(#3-10) Open the math 140 fall 2015 survey data at www.matt-teachout.org. Copy and paste the smoking status column and the type of transportation column next to each other in a new excel spread sheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for smoking status and transportation. Use the table to answer the following questions.

3. What percent of the math 140 students smoke?
4. What proportion of the math 140 students drive alone to school?
5. What percent of the math 140 students both carpool and do not smoke?
6. What proportion of the math 140 students both smoke and drive alone to school?
7. What percent of the math 140 students either do not smoke or are dropped off by someone?
8. What proportion of the math 140 students either walk to school or smoke?
9. What percent of the smoking math 140 students carpool? What percent of the non-smoking math 140 students carpool? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between smoking and carpooling to school?
10. What proportion of the drive alone math 140 students smoke? What proportion of the dropped off math 140 students smoke? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between smoking and the type of transportation?

(#11-18) Open the math 140 fall 2015 survey data at www.matt-teachout.org. Copy and paste the texting and driving column and the car accident column next to each other in a new excel spreadsheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for texting and driving and car accidents. Use the table to answer the following questions.

11. What percent of the math 140 students text and drive?



12. What proportion of the math 140 students have been in a car accident?
13. What percent of the math 140 students both text and drive and have been in a car accident?
14. What proportion of the math 140 students do not text and drive and have not been in a car accident?
15. What percent of the math 140 students either text and drive or have not been in a car accident?
16. What proportion of the math 140 students either do not text and drive or have been in a car accident?
17. What percent of the text and drive math 140 students have been in a car accident? What percent of the not text and drive math 140 students have been in a car accident? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between texting and driving and car accidents?
18. What proportion of the car accident math 140 students text and drive? What proportion of the no car accident math 140 students text and drive? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between texting and driving and car accidents?

(#19-26) Open the math 140 fall 2015 survey data at www.matt-teachout.org. Copy and paste the tattoos column and the favorite social media column next to each other in a new excel spread sheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for tattoos and favorite social media. Use the table to answer the following questions.

19. What percent of the math 140 students have a tattoo?
20. What proportion of the math 140 students prefer snapchat?
21. What percent of the math 140 students both prefer Facebook and do not have a tattoo?
22. What proportion of the math 140 students both have a tattoo and prefer twitter?
23. What percent of the math 140 students either prefer Instagram or have a tattoo?
24. What proportion of the math 140 students either prefer twitter or do not have a tattoo?
25. What percent of the tattoo math 140 students prefer twitter? What percent of the no tattoo math 140 students prefer twitter? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between liking twitter and having a tattoo?
26. What proportion of the Instagram math 140 students have a tattoo? What proportion of the Facebook math 140 students have a tattoo? Do the proportions appear to be close or significantly different? What does this tell us about the relationship between social media and having a tattoo?

(#27-34) Open the car data at www.matt-teachout.org. Copy and paste the country column and the cylinders column next to each other in a new excel spread sheet. Then copy both columns together. Open StatKey at www.lock5stat.com. Under the "Descriptive Statistics and Graphs" menu, click on "Two Categorical Variables". Paste the two columns into StatKey. Be sure to check the boxes for "Raw Data" and "Header Row" and push "OK". Use StatKey to create a contingency table for the country and cylinders. Use the table to answer the following questions.

27. What percent of the cars were made in Germany?
28. What proportion of the cars have six cylinders?
29. What percent of the cars have four cylinders and are made in Japan?
30. What proportion of the cars have eight cylinders and are made in the U.S.?
31. What percent of the cars either have six cylinders or are made in Germany?
32. What proportion of the cars either have six cylinders or are made in the Japan?



33. What proportion of the cars made in Japan have four cylinders? What proportion of cars made in Germany have four cylinders? Are the proportions appear to be close or significantly different? What does this tell us about the relationship between the country and the number of cylinders?

34. What proportion of the cars with six cylinders were made in the U.S.A? What proportion of the cars with eight cylinders were made in the U.S.A? Are the proportions appear to be close or significantly different? What does this tell us about the relationship between cars made in the U.S.A and the number of cylinders?

Section 4F – Categorical Relationships: Categorical Association Test

In this chapter, we looked at the Goodness of Fit test. The Goodness of Fit Test determines if a single proportion is related to some other categorical variable. In this section, we will look at situations where more than one proportion is involved. When we have multiple different proportions in multiple groups, we call this the Categorical Association Test.

In the Categorical Association Test, we will be determining if categorical variables are related or not. Many students confuse the Goodness of Fit test and the Categorical Association Test because they are both categorical relationship tests. Look at your sample counts. If you have a single observed count for each group, you are doing a Goodness of Fit test since we are only looking at one proportion in the groups. If your observed counts are summarized in a contingency table, then more than one proportion is involved in your groups. That makes it a Categorical Association Test.

For example, suppose we wanted to see if the amount of education a person has is related to their health. Notice the amount of education has multiple options and the health status has multiple options. This cannot be a Goodness of Fit test. The observed counts are summarized in a contingency (two-way) table so we will be using the Categorical Association Test.

	Excellent Health	Good Health	Fair Health	Poor Health
Less than High School	72	202	199	62
High School Diploma	465	877	358	108
Some College/Associates Degree	80	138	49	11
Bachelor's Degree	229	276	64	12
Graduate Degree	130	147	32	2

A Goodness of Fit Test would only look at one of these. For example, suppose we want to see if the proportion for excellent health is related to education. In that case, the data would look like this.

	Excellent Health
Less than High School	72
High School Diploma	465
Some College/Associates Degree	80
Bachelor's Degree	229
Graduate Degree	130



The Categorical Association Test

So let us look at the categorical association test. This test determines if categories are related or not. The categories can have multiple options. The sample data for this test is either two raw categorical data sets or summary counts summarized in a contingency (two-way) table.

Null and Alternative Hypothesis

In the last section, we examined conditional proportions. We saw that we need to compare conditional proportions from the same variable. If the conditional proportions are equal or close in our groups, it indicates that the categorical variables are not related. If the conditional proportions are significantly different, then the categorical variables are related.

For example, we will want to compare the proportions for excellent health in all of our education groups. We will also want to compare the proportions for poor health in all of our education groups and so on. We will not want to compare the proportion of excellent health to poor health since they are not the same variable.

	Excellent Health	Good Health	Fair Health	Poor Health
Less than High School	72	202	199	62
High School Diploma	465	877	358	108
Some College/Associates Degree	80	138	49	11
Bachelor's Degree	229	276	64	12
Graduate Degree	130	147	32	2

If there is no relationship between the categories, we expect the conditional proportions for each variable to be equal in all the groups. If there is a relationship between the categories, we expect at least one or more of the conditional proportions for each variable to be different. It is difficult to specify all of the conditional proportions and groups, so to summarize, we often say that the “distribution of conditional proportions are the same” or the “distribution of conditional proportions are different”.

Note

- Saying that the categories are “related” can also be described as “associated” or “dependent”.
- Saying that the categories are “not related” can also be described as “not associated” or “independent”.

Categorical Association Test Null and Alternative Hypothesis

H_0 : The categories are not related (distribution of conditional proportions are equal)

H_A : The categories are related (distribution of conditional proportion are different)

The Chi-Squared Test Statistic (χ^2)

Since multiple proportions in multiple groups are involved, we will be using the Chi-Squared test statistic (χ^2) again. In our previous study of using the Chi-squared test statistic, we saw that this test statistic compares the observed sample counts to the expected counts based on the null hypothesis.

$$\text{Chi-Squared Test Statistic } (\chi^2) = \sum \frac{(O-E)^2}{E}$$



The expected counts are what we expect to happen if the null hypothesis is true. If the null hypothesis is true and there is no relationship between the categories, we expect the conditional proportions to be equal in the various groups. If we multiply the equal proportions by the size of the group, we can get our expected counts. Here is a formula that computer programs use to calculate the expected counts.

$$\text{Expected Counts (for contingency table)} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

How does this formula give us the expected counts and account for equal conditional proportions? To answer this, we will look at an example.

Example 1

A sample of 75 people were paid to participate in an experiment. The goal of the experiment was to determine if listening to music is related to a person's ability to memorize information. The people were randomly assigned into three groups. One group tried to memorize some information while listening to their favorite music. Another group tried to memorize some information while listening to music they hated. The third group tried to memorize some information in a silent room. All of the people attempted to memorize the same information and took a test to determine how much of the information they remembered. Confounding variables were controlled. For example, the volume of music was the same in the music groups. We will use a 10% significance level.

	Liked Music	Disliked Music	No Music	Total
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	Grand Total = 75

H_0 : Listening to music and memorizing information are not related (not associated, independent).

H_A : Listening to music and memorizing information are related (associated, dependent). CLAIM

Remember conditional proportions are important to explore when analyzing relationships between categorical variables. Here are two very important principles.

1. When conditional proportions were close or equal, it indicated that the variables were not related to each other.
2. When conditional proportions were significantly different, it indicated that the variables were related to each other.

Let us look at some conditional proportions. Remember we need to compare the same variable proportion in different groups.

Let us calculate the proportion of people in the liked music group that were able to memorize a lot of the information? Let us compare that to the proportion of people in the hated music group that were able to memorize a lot of the information.

$$P(\text{high retention} \mid \text{liked music}) = 10/24 = 0.417 \text{ or } 41.7\%$$

$$P(\text{high retention} \mid \text{disliked music}) = 11/26 = 0.423 \text{ or } 42.3\%$$

These two conditional probabilities are close and so indicate that the music and high retention are not related or independent.

Here lies the fundamental problem. We are not really taking the entire contingency (two-way) table and all of the conditional probabilities into account. If we look at another conditional probability, we may come to a different conclusion. Look at these two.



Let us calculate the proportion of people in the liked music group that were able to memorize a lot of the information? Let us compare that to the proportion of people in the no music group that were able to memorize a lot of the information.

$$P(\text{high retention} \mid \text{liked music}) = 10/24 = 0.417 \text{ or } 41.7\%$$

$$P(\text{high retention} \mid \text{no music}) = 18/25 = 72\%$$

These two probabilities are significantly different and so indicate that the music and high retention are related.

So it is difficult to determine if categorical variables are related or not by just looking at two conditional proportions. We need a better way to do this.

Calculating the Chi-Squared Test Statistic

A much better way to determine if the categories are related or not is by using the chi-squared test statistic. It takes into account all of the conditional probabilities possible instead of relying on only two. Remember to calculate the chi-squared test statistic, we need to compare the expected counts (expected frequencies) to the observed counts (observed frequencies).

Expected Counts

The “expected counts” or “expected frequencies” are what we expect to happen if the null hypothesis is true. For the Categorical Association Test, the null hypothesis is that the categories are not related (independent). This would imply that the distribution of conditional proportions are equal.

Let us work this out for the music and retention problem.

	Liked Music	Disliked Music	No Music	Total
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	Grand Total = 75

If the null hypothesis is true, we expect the proportion for high retention to be the same regardless of the music choice. If we disregard music, then the proportion of high retention would be the amount of high retention (39) divided by the grand total (75). So if the null hypothesis is true and music is not related to retention, then 52% of every group should memorize a significant about of the information.

$$P(\text{high retention}) = 39/75 = 0.52$$

Remember the expected values are found by multiplying the proportion times the total number of people or objects in that group.

$$E = n \times p$$

Only the n is not the grand total, it is the total for each column (each music group).

If the null hypothesis is true, we expect the p for high retention to always be 0.52 and the expected values will be 0.52 x total students for each music choice.

$$E_{\text{liked music high retention}} = n \times p = 24 \times 0.52 = 12.48$$

$$E_{\text{hated music high retention}} = n \times p = 26 \times 0.52 = 13.52$$

$$E_{\text{no music high retention}} = n \times p = 25 \times 0.52 = 13.0$$



Similarly, we expect the proportion for low retention to be the same in all of the music groups. If we disregard music, then the proportion of low retention would be the amount of low retention (36) divided by the grand total (75). So if the null hypothesis is true and music is not related to retention, then 48% of every group should not be able to memorize much of the information.

$$P(\text{low retention}) = 36/75 = 0.48$$

So if the null is true we expect the p for low retention to always be 0.48 and the expected counts will be $0.48 \times$ total people for each music group.

$$E_{\text{liked music low retention}} = n \times p = 24 \times 0.48 = 11.52$$

$$E_{\text{hated music low retention}} = n \times p = 26 \times 0.48 = 12.48$$

$$E_{\text{no music low retention}} = n \times p = 25 \times 0.48 = 12.0$$

Earlier we saw that computer programs often use this formula to calculate the expected counts.

$$\text{Expected Counts (for contingency table)} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

Notice that the column total is the total number of people in each music group. The row total divided by the grand total is the proportion that must be equal for all of the groups.

$$E = n \times p = \text{Column Total} \times \left(\frac{\text{Row Total}}{\text{Grand Total}} \right)$$

Now let us calculate the Chi-Squared Test Statistic

We learned that the Chi-Squared test statistic is a comparison of the observed sample values and the expected values from the null hypothesis. Here is the formula again.

$$\text{Chi-Squared Test Statistic } (\chi^2) = \sum \frac{(O-E)^2}{E}$$

So Chi-Squared subtracts the observed and expected values to find the difference. Since some differences are negative, it squares the differences. It also divides by E to make it a kind of average of squares and finally it adds up these values for every variable.

Here is the sentence to explain Chi-Squared again:

“The sum of the averages of the squares of the differences between the observed sample data and the expected values if the null hypothesis were true.”

	Liked Music	Disliked Music	No Music	Total
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	Grand Total = 75

In this example, the numbers in the two-way table are the observed counts. Note: The observed counts do not include the totals! This two-way table has two rows and three columns (not counting totals). This is often called a “two by three” (2x3) table. So we have six observed counts and six expected counts.

	Liked Music	Disliked Music	No Music
High Retention	10	11	18
Low Retention	14	15	7

Let us calculate the Chi-Squared test statistic for this problem. Here are the expected counts again. It is good to label so that you subtract the correct expected count from the correct observed count.



$$E_{\text{liked music high retention}} = n \times p = 24 \times 0.52 = 12.48$$

$$E_{\text{hated music high retention}} = n \times p = 26 \times 0.52 = 13.52$$

$$E_{\text{no music high retention}} = n \times p = 25 \times 0.52 = 13.0$$

$$E_{\text{liked music low retention}} = n \times p = 24 \times 0.48 = 11.52$$

$$E_{\text{hated music low retention}} = n \times p = 26 \times 0.48 = 12.48$$

$$E_{\text{no music low retention}} = n \times p = 25 \times 0.48 = 12.0$$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(10-12.48)^2}{12.48} + \frac{(11-13.52)^2}{13.52} + \frac{(18-13)^2}{13} + \frac{(14-11.52)^2}{11.52} + \frac{(15-12.48)^2}{12.48} + \frac{(7-12)^2}{12}$$

$$\approx 0.49282 + 0.46970 + 1.92308 + 0.53388 + 0.50885 + 2.08333 \approx 6.012$$

The numbers that were added to get the Chi-Squared test statistic are called the “Contributions to Chi-Squared”. Notice that the largest contributions to Chi-squared were 1.92308 and 2.08333. These calculations came from the low and high retention from the no music group.

Is this Chi-squared test statistic significant?

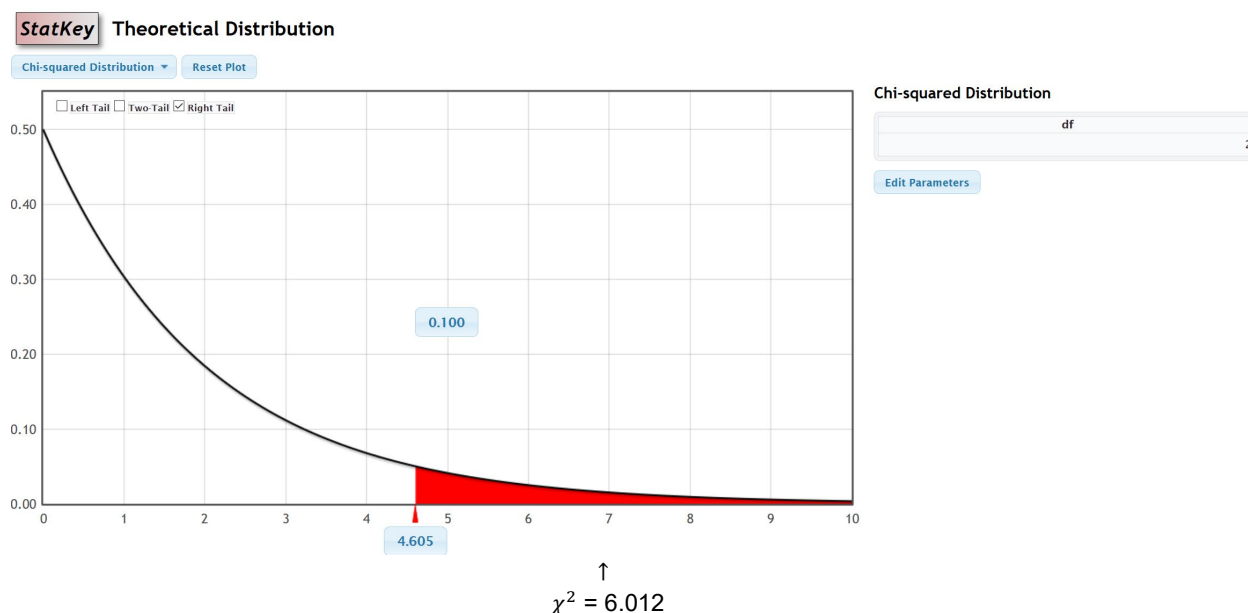
In a contingency table the degrees of freedom is the number of rows minus one times the number of columns minus one. Note: Do not include the totals when you count the number of rows and columns.

Degrees of Freedom (for Contingency Table Data) = $(r - 1) \times (c - 1)$
where “r” is the number of rows and “c” is the number of columns.

In the music and retention data, there were two rows and three columns so the degrees of freedom will be two.

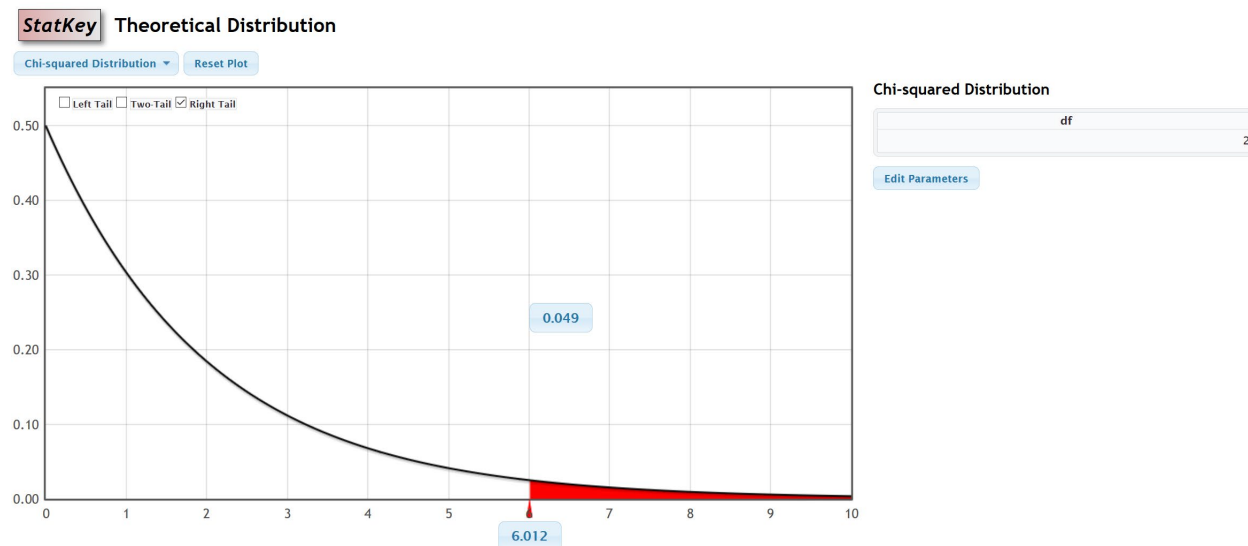
$$\text{Degrees of Freedom (for Contingency Table Data)} = (r - 1) \times (c - 1) = (2 - 1) \times (3 - 1) = 1 \times 2 = 2$$

The Categorical Association Test is a right tailed test. We can use the degrees of freedom to look up the critical value using the Chi-Squared theoretical distribution calculator in StatKey. Go to www.lock5stat.com and click on StatKey. Under the “Theoretical Distributions” menu, click on “ χ^2 ”. Put in the degrees of freedom and click “right-tail”. Since we are using a 10% significance level, we will enter 0.10 in the proportion for the right tail.



Notice that the critical value was 4.605. So our test statistic must be 4.605 or greater to be considered significant. Our Chi-squared test statistic was 6.012, which is in the right tail. This tells us that the sample data significantly disagrees with the null hypothesis. It also tells us that our observed counts are significantly different from our expected counts.

We can also use the theoretical Chi-squared curve to calculate the P-value. Just put the test statistic of 6.012 in the bottom box. Notice the computer calculated a P-value of 0.049 or 4.9%. This is less than our 10% significance level.



We can also use a randomized simulation in StatKey. Under the “More Advanced Randomization Tests” menu, click on “ χ^2 Test for Association”. We will need to type in our observed counts into StatKey. Do not include the totals. The computer will calculate the totals automatically.

	Liked Music	Disliked Music	No Music
High Retention	10	11	18
Low Retention	14	15	7

Under the “Edit Data” menu, type in the contingency table with commas. Notice that “[blank]” must be in the top left corner.

[blank], Liked Music, Disliked Music, No Music

High Retention, 10, 11, 18

Low Retention, 14, 15, 7



Edit data
✕

[blank], Liked Music, Disliked Music, No Music
High Retention, 10, 11, 18
Low Retention, 14, 15, 7

☐ Raw Data
☒ Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only two columns if it contains raw data. Summary counts tables require both row and column headers.

Ok

Under “Original Sample”, we see that StatKey has calculated the Chi-Squared test statistic of 6.012. If you click on “Show Details”, you can see the expected counts and the contributions to chi-squared.

Original Sample Show Details

$n = 75, \chi^2 = 6.012$

	Liked Music	Disliked Music	No Music	Total
High Retention	10	11	18	39
Low Retention	14	15	7	36
Total	24	26	25	75

Detailed Sample Table ✕

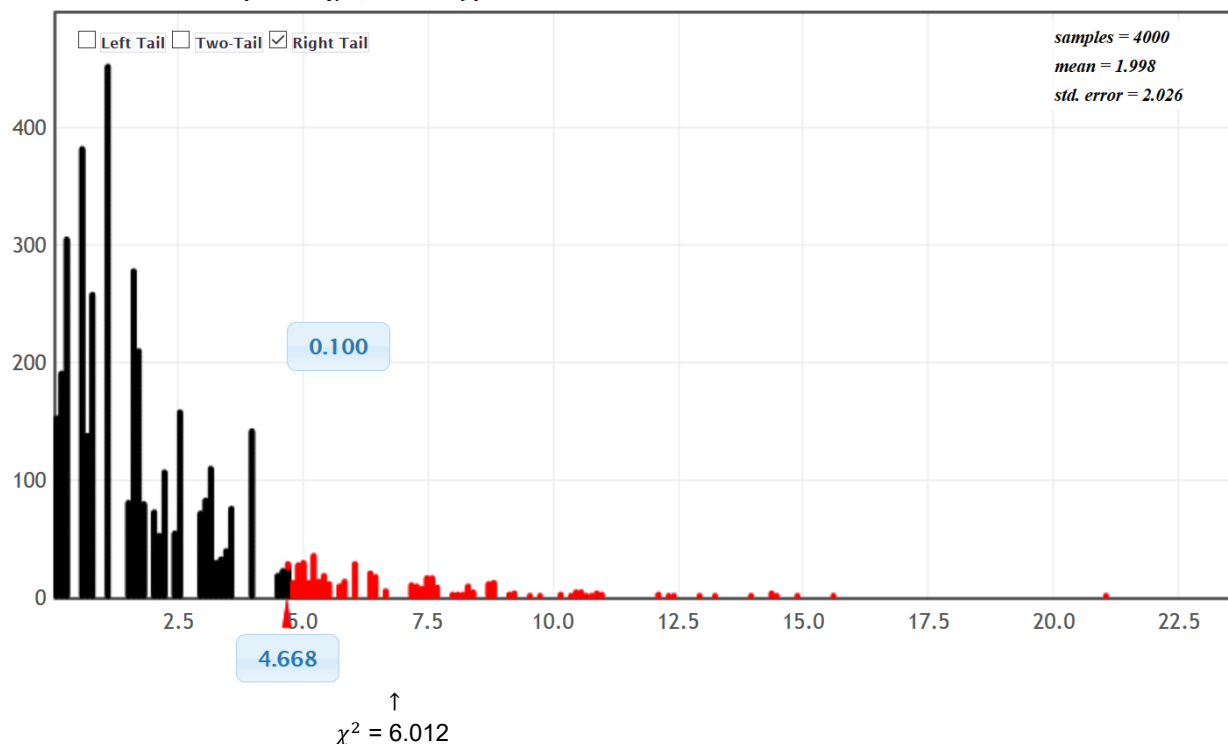
	Liked Music	Disliked Music	No Music	Total
High Retention	10 12.5 0.493	11 13.5 0.47	18 13 1.923	39
Low Retention	14 11.5 0.534	15 12.5 0.509	7 12 2.083	36
Total	24	26	25	75

Observed, Expected, Contribution to χ^2



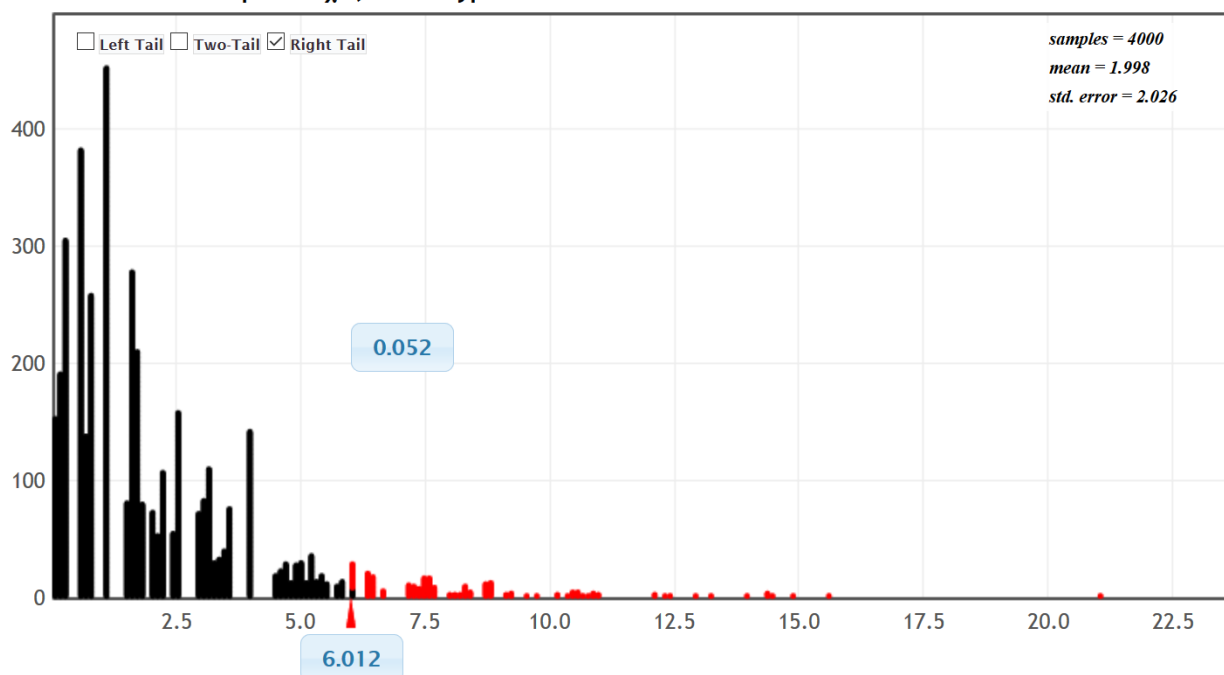
If we click “Generate 1000 Samples” a few times, we get our randomized simulation. Notice the null hypothesis is “No Association” (not related). Putting our 10% significance level in the tail proportion, gives us an approximate critical value of 4.668. This is close to what we got with the theoretical curve. Notice that our Chi-squared test statistic of 6.012 does fall in the tail of the simulation.

Randomization Dotplot of χ^2 , Null hypothesis: No Association



We can also use the simulation to calculate the approximate P-value. Just put the test statistic of 6.012 into the bottom box. Notice the P-value came out to be 0.052. This is almost the same P-value as we calculated with the theoretical Chi-squared distribution.



Randomization Dotplot of χ^2 , Null hypothesis: No Association

Calculating the Chi-Squared test statist and P-value with Statcato

- First type in the contingency (two-way) table exactly as you see it into Statcato. The column titles will be in the grey cells (VAR) at the top, but the row titles will be in the regular cells. Do not type the totals. Statcato will automatically calculate them.

	C1	C2	C3	C4
Var	High Retention	Liked Music	Disliked Music	No Music
1		10	11	18
2		14	15	7

- Go to the “statistics” menu and click on “multinomial experiment” then “chi-square contingency table”. Hold the control key down and click on the columns that have your observed frequencies (C2, C3 & C4). Do not include the row labels (C1). They can be added later.

Chi-Square Test: Contingency Table:

	C2 Liked Music	C3 Disliked Music	C4 No Music	Total
High Retention	10.0 (12.48) [0.49]	11.0 (13.52) [0.47]	18.0 (13.0) [1.92]	39.0
Low Retention	14.0 (11.52) [0.53]	15.0 (12.48) [0.51]	7.0 (12.0) [2.08]	36.0
Total	24.0	26.0	25.0	75.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	2	6.0117	5.9915	0.0495



Notice the test statistic, critical value and P-value in Statcato are about the same as we got in the simulation.

The expected counts (expected frequencies) are in parenthesis in the Statcato printout and the contributions to chi-squared are in brackets.

Assumptions

The assumptions for the Categorical Association Test are as follows:

Categorical Association Test Assumptions

- The categorical sample or samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- If multiple samples are collected, then the data values between the samples should be independent of each other.
- The expected counts from the null hypothesis should be at least five.

Did the music and retention data meet the assumptions?

Random? Yes. This was an experiment and the groups were randomly assigned.

Were the expected counts at least 5? Yes. The expected counts were 12.4, 13.52, 13.0, 11.52, 12.48, and 12.0. All the expected frequencies were at least five.

Independence? The individuals in the experiment were randomly assigned, so we can probably assume the data met the independence requirement.

Writing the conclusion

Should we reject the null hypothesis or fail to reject the null hypothesis? Since our P-value is lower than our significance level, we reject the null hypothesis.

Remember a conclusion must address the evidence and the claim. Since our P-value is low, we have significant evidence. We rejected the null hypothesis, but the claim was the alternative hypothesis. We will therefore support the claim.

H_0 : Listening to music and memorizing information are not related (not associated, independent).

H_A : Listening to music and memorizing information are related (associated, dependent). CLAIM

Conclusion: We have significant sample evidence to support the claim that listening to music and retaining information are related.

In fact, since confounding variables were controlled, the experiment proves cause and effect. The “no music” group did significantly better than either of the music groups. This experiment appears to indicate that when a person listens to either music they like or music they hate, they have a harder time memorizing information.

Note about “Independence” and “Homogeneity”

Sometimes we obtain multiple categorical data from one random sample of people or objects. We ask multiple categorical questions from the same group of people. Statisticians sometimes refer to that situation as an “independence” test.

If we collect the same categorical data from multiple random samples, then that is sometimes referred to as a “homogeneity” test. We may ask the same categorical question from different groups of people.

Whether you collected the data from one sample or multiple samples, the data still summarizes into a contingency table. Look at the following sample data.



	Business	English	History	Music	Biology	Math
Female	89	71	62	48	56	9
Male	112	58	59	53	62	13

Suppose we took one random sample of college students and asked them two categorical questions. What gender do you most identify with? What is your major? This would be referred to as an “independence” test.

We could collect this data in another way. Suppose we took a random sample of female college students and asked them what their major was. Later we took a random sample of male college students and asked them the same question. This would now be referred to as a “homogeneity” test.

Notes about the Categorical Association Test

- The Categorical Association Test is used to determine if categories with multiple options are related or not.
- The categorical association test is always a right-tailed test.
- The degrees of freedom for the categorical association test is the number of rows minus one times the number of columns minus one. $df = (r - 1) \times (c - 1)$
- The categorical association test uses the Chi-squared test statistic (χ^2), which compares the observed sample counts to the expected counts if the null hypothesis was true.
- Always use a computer to calculate the test statistic. Focus on being able to interpret and judge significance.

Practice Problems Section 4F

(#1-10) Use each of the following categorical association χ^2 -test statistics and the corresponding critical values to fill out the table.

	χ^2 -test stat	Sentence to explain χ^2 -test statistic.	Critical Value	Does the χ^2 -test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	+1.573		+4.117		
2.	+6.226		+5.118		
3.	+2.144		+4.121		
4.	+3.415		+5.091		
5.	+13.718		+7.189		
6.	+0.972		+4.812		
7.	+31.652		+12.557		
8.	+11.185		+5.181		
9.	+25.443		+7.008		
10.	+1.133		+8.336		

(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.263			10%			
12.	0.0042			1%			
13.	5.22×10^{-4}			5%			
14.	0.0639			1%			
15.	0			5%			
16.	0.539			10%			



17.	0.0419			5%			
18.	0.0027			10%			
19.	7.73×10^{-8}			1%			
20.	0.674			5%			

21. If we have two raw categorical data sets, what must we click on in Statcato to perform a categorical association test?

22. If we have summary counts organized in a contingency table, what must we click on in Statcato to perform a categorical association test?

23. What are the assumptions for a categorical association test if the data was collected from on random sample?

24. What are the assumptions for a categorical association test if the data was collected from multiple random samples?

25. How are the expected counts calculated in a categorical association test?

26. If the expected counts from the null hypothesis are significantly different from the observed sample counts, describe the effect on the Chi-Squared test statistic.

27. If the expected counts from the null hypothesis are close to the observed sample counts, describe the effect on the Chi-Squared test statistic.

(#28-31) *Directions: For each of the following problems, use the Statcato printout provided to answer the following questions.*

a) *Write the null and alternative hypothesis. Make sure to label which one is the claim.*

b) *Check the assumptions for the categorical association test.*

c) *What is the Chi-squared test statistic? Write a sentence to explain the test statistic.*

d) *Does the test statistic fall in the tail determined by the critical value?*

e) *Does the sample data significantly disagree with the null hypothesis? Explain your answer.*

f) *Are the observed counts significantly different from the expected counts? Explain your answer.*

g) *What is the P-value? Write a sentence to explain the P-value.*

h) *Compare the P-value to the significance level. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.*

i) *If the null hypothesis was true, could the sample data or more extreme have occurred by sampling variability or is it unlikely to be sampling variability? Explain your answer.*

j) *Write a conclusion for the test addressing evidence and the claim. Explain your conclusion in non-technical language.*

k) *Are the categories related or not? Explain your answer.*



28. A random sample of male college students were asked their major. Later, a random sample of female college students were asked their major. The goal of the study was to show that gender is not related to major. Use a 5% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Business	English	History	Music	Biology	Math	Total
Female	89.0 (97.30) [0.71]	71.0 (62.45) [1.17]	62.0 (58.58) [0.20]	48.0 (48.89) [0.02]	56.0 (57.12) [0.02]	9.0 (10.65) [0.26]	335.0
Male	112.0 (103.70) [0.67]	58.0 (66.55) [1.10]	59.0 (62.42) [0.19]	53.0 (52.11) [0.02]	62.0 (60.88) [0.02]	13.0 (11.35) [0.24]	357.0
Total	201.0	129.0	121.0	101.0	118.0	22.0	692.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	5	4.6014	11.0705	0.4664

29. A random sample of adults were asked their blood type and Rh status. (Blood tests were provided for those that did not know their blood type and Rh status.) The goal of the study was to show that blood type is related to Rh status (dependent). Use a 10% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Type A	Type B	Type AB	Type O	Total
Rh+	35.0 (36.03) [0.03]	24.0 (23.0) [0.04]	11.0 (16.1) [1.62]	91.0 (85.87) [0.31]	161.0
Rh-	12.0 (10.97) [0.10]	6.0 (7.0) [0.14]	10.0 (4.9) [5.31]	21.0 (26.13) [1.01]	49.0
Total	47.0	30.0	21.0	112.0	210.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.10	3	8.5522	6.2514	0.0359



30. A hospital wanted to determine if the age of a patient is not related to what part of the hospital they were in. They took a random sample of patients that have visited their hospital and determined both their age and the part of the hospital. The ages were broken up into age groups. Use a 1% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Med/Surg	ICU	SDS	ER	Total
18-35 years old	19.0 (19.12) [7.98 · 10 ⁻⁴]	4.0 (11.47) [4.87]	25.0 (17.85) [2.87]	16.0 (15.55) [0.01]	64.0
36-49 years old	27.0 (19.42) [2.96]	7.0 (11.65) [1.86]	22.0 (18.13) [0.83]	9.0 (15.80) [2.92]	65.0
50-64 years old	17.0 (18.53) [0.13]	13.0 (11.12) [0.32]	15.0 (17.29) [0.30]	17.0 (15.07) [0.25]	62.0
65+ years old	12.0 (17.93) [1.96]	21.0 (10.76) [9.75]	8.0 (16.73) [4.56]	19.0 (14.58) [1.34]	60.0
Total	75.0	45.0	70.0	61.0	251.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.01	9	34.9208	21.666	6.153 · 10 ⁻⁵

31. A random sample of American adults was taken and their health and education status obtained. Test to test the claim that health and education are related. Use a 5% significance level and the Statcato printout below to answer the questions given above.

Chi-Square Test: Contingency Table:

	Excellent Health	Good Health	Fair Health	Poor Health	Total
Less Than High School	72.0 (148.64) [39.51]	202.0 (249.76) [9.13]	199.0 (106.91) [79.33]	62.0 (29.70) [35.14]	535.0
High School Diploma	465.0 (502.31) [2.77]	877.0 (844.04) [1.29]	358.0 (361.29) [0.03]	108.0 (100.36) [0.58]	1808.0
Some College / Associates Degree	80.0 (77.24) [0.10]	138.0 (129.78) [0.52]	49.0 (55.55) [0.77]	11.0 (15.43) [1.27]	278.0
Bachelor's Degree	229.0 (161.42) [28.30]	276.0 (271.23) [0.08]	64.0 (116.10) [23.38]	12.0 (32.25) [12.72]	581.0
Graduate Degree	130.0 (86.40) [22.00]	147.0 (145.19) [0.02]	32.0 (62.15) [14.62]	2.0 (17.26) [13.49]	311.0
Total	976.0	1640.0	702.0	195.0	3513.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	12	285.0610	21.0261	0



(#32-35) Directions: Use StatKey at www.lock5stat.com to simulate the following chi-squared categorical association tests. Go to the “More Advanced Randomization Tests” menu at the bottom of the StatKey page. Click on the button that says, “ χ^2 Test for Association”. Click on “Edit Data” and type in the contingency table provided. Click on “Generate 1000 Samples” a few times to create the simulated sampling distribution and answer the following questions.

- a) Write the null and alternative hypothesis. Make sure to label which one is the claim.
- b) Check the assumptions for the categorical association test. Assume the data was collected randomly. Under “Original Sample”, click on “Show Details” to see the expected counts.
- c) Use the formula $df = (r - 1)(c - 1)$ to calculate the degrees of freedom. “r” is the number of rows and “c” is the number of columns not counting the totals.
- d) What is the Chi-squared test statistic? Write a sentence to explain the test statistic.
- e) Put the significance level proportion in the right tail proportion to calculate the critical value. What is the critical value? (Answers will vary slightly.) Does the original sample χ^2 test statistic fall in the tail determined by the critical value?
- f) Does the sample data significantly disagree with the null hypothesis? Explain your answer.
- g) Are the observed counts significantly different from the expected counts? Explain your answer.
- h) Put the original sample test χ^2 test statistic in the bottom box in the simulation to calculate the P-value. What is the P-value? (Answers will vary slightly.) Write a sentence to explain the P-value.
- i) Compare the P-value to the significance level. Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- j) If the null hypothesis was true, could the sample data or more extreme have occurred by sampling variability or is it unlikely to be sampling variability? Explain your answer.
- k) Write a conclusion for the test addressing evidence and the claim. Explain your conclusion in non-technical language.
- l) Are the categories related or not? Explain your answer.

32. We want to know if the state a home is built in is related to the size of the home. A random sample of homes in the U.S was taken. Click on “Edit Data” in StatKey and type in the following contingency table. Do not forget to include a space after the commas. Use a 5% significance level and randomized simulation to test the claim that the state is not related to size of the home.

[blank], CA, NJ, NY, PA

Large, 7, 6, 7, 3

Small, 23, 24, 23, 27

33. Open the “Car Data” at www.matt-teachout.org. Copy and paste the “Country” and “Cylinders” columns next to each other in a new Excel spreadsheet. Then copy the two columns together. Click on “Edit Data” in StatKey and paste the two columns into StatKey. Use a 1% significance level to test the claim that the country a car is made in is related to the cylinders. Answer the questions above.



34. We want to show that gender is related to getting an award. A random sample of people that won famous awards in the Olympic, Academia, and Nobel was taken and their gender was noted. Click on “Edit Data” in StatKey and type in the following contingency table. Do not forget to include a space after the commas. Use a 10% significance level and randomized simulation to test the claim that awards are related to gender.

[blank], Olympic, Academy, Nobel

Male, 109, 11, 73

Female, 73, 20, 76

35. Open the “Math 140 Fall 2015 Survey Data” at www.matt-teachout.org. Copy and paste the “Tattoo” and “Favorite Social Media” columns next to each other in a new Excel spreadsheet. Then copy the two columns together. Click on “Edit Data” in StatKey and paste the two columns into StatKey. Use a 5% significance level to test the claim that having a tattoo or not is not related to social media. Answer the questions above.

Section 4G – Quantitative Relationships: Correlation and Regression

Vocabulary

Correlation: Statistical analysis that determines if there is a relationship between two different quantitative variables.

Regression: Statistical analysis that involves finding the line or model that best fits a quantitative relationship, using the model to make predictions, and analyzing error in those predictions.

Explanatory Variable (x): Another name for the x-variable or independent variable in a correlation study.

Response Variable (y): Another name for the y-variable or dependent variable in a correlation study.

Correlation Coefficient (r): A statistic between -1 and $+1$ that measures the strength and direction of linear relationships between two quantitative variables.

R-squared (r^2): Also called the coefficient of determination. This statistic measures the percent of variability in the y-variable that can be explained by the linear relationship with the x-variable.

Residual ($y - \hat{y}$): The vertical distance between the regression line and a point in the scatterplot.

Standard Deviation of the Residual Errors (s_e): A statistic that measures how far points in a scatterplot are from the regression line on average and measures the average amount of prediction error.

Slope (b_1): The amount of increase or decrease in the y-variable for every one-unit increase in the x-variable.

Y-Intercept (b_0): The predicted y-value when the x-value is zero.

Regression Line ($\hat{y} = b_0 + b_1x$): Also called the line of best fit or the line of least squares. This line minimizes the vertical distances between it and all the points in the scatterplot.

Scatterplot: A graph for visualizing the relationship between two quantitative ordered pair variables. The ordered pairs (x, y) are plotted on the rectangular coordinate system.

Residual Plot: A graph that pairs the residuals with the x values. This graph should be evenly spread out and not fan shaped.

Histogram of the Residuals: A graph showing the shape of the residuals. This graph should be nearly normal and centered close to zero.



Introduction

Sometimes we want to know if two different quantitative variables are related to each other. This kind of relationship study is difficult because the units are different. We cannot directly compare the height of man in inches to his weight in pounds. Inches and pounds are completely different. Statisticians and mathematicians developed a type of analysis for this situation called “correlation and regression”. The idea is to let one variable be X and the other variable be Y . Then use ordered pair data to create a graph called a scatterplot and look for patterns. The most common is a linear pattern (correlation). If we see a linear pattern, we can also calculate the line that best fits the data and use this line to make predictions (regression).

Choosing your variables

It is important to determine which variable will be X and which variable will be Y . In statistics, we call the X -variable the “explanatory variable” or the “independent variable”. We call the Y -variable the “response variable” or “dependent variable”. How do we choose? Here are a couple key questions to ask yourself.

- Does one variable respond more than the other does?
- Which variable is the focus of the study and the variable I might want to make predictions about?

Let us look at some examples.

Example: Year (time) and unemployment rates in U.S.

Ask yourself the following question. Does one of the variables respond more than the other? Does time fluctuate in response to the unemployment rate? That does not sound right. Time seems to go on no matter what happens with unemployment. Do you think unemployment might fluctuate in response to time? That seems more likely. So we should let the explanatory variable X be time (years) and let the response variable y be unemployment rate. Unemployment responds to time, but not the other way around.

Example: The unemployment rate in U.S. and the national debt in the U.S.

These variables respond to each other, so either variable could be the response variable Y . In this case, pick the response variable (Y) to be the one you are most interested in (focus of the study) or the variable you may want to make predictions about. If there is a relationship, then the Y -variable will be the variable you can make predictions about.

Suppose the focus of your study and the variable you want to predict is the national debt. Unemployment may just be one factor that may be related to the national debt. If that is the case, you should make the national debt your response variable Y . By default, that means that unemployment rate would be explanatory variable X .

Correlation Graphs and Statistics with StatKey

To study the relationship between two different quantitative variable, you will need ordered pair data. For example, we will need the height and weight of the same men, or the unemployment rate and national debt of the same countries. Decide which variable should be X and which variable should be Y . The computer will then make ordered pairs from your data (X, Y) and plot all the points on the rectangular coordinate system. This graph of all the ordered pairs is called a scatterplot.

Example

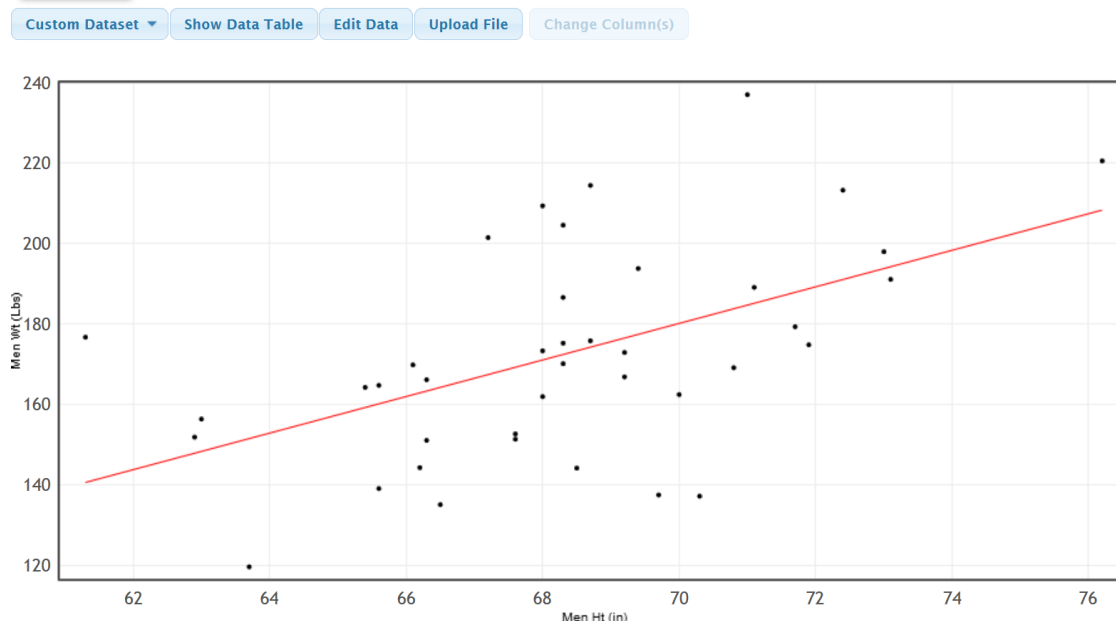
Suppose we want to study if the weights in pounds of the men in the health data is related to their heights in inches. I am most interested in predicting the weights of men from their heights so I will let the weight be the response variable Y and height be the explanatory variable X . Notice these are ordered pairs, since the heights and weights came from the same 40 men.



To put the data into StatKey, you will want to open a fresh excel spreadsheet and place the two data sets side by side. These two data sets are already next to each other in the health data, but in general, the data sets may not be. Copy the two columns of data together.

Go to www.lock5stat.com and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “two quantitative variables”. Under the “edit data” tab, paste the height and weight data into StatKey. The graph you see is the scatterplot. Notice StatKey has placed the heights on the horizontal x-axis and the weights on the vertical y-axis. If it is backward, simply click the “switch variables” button. It is also nice to check the “show regression line” box. The regression line is the line that best fits the points in the scatterplot. StatKey has also given us some statistics to help understand the relationship.

StatKey Descriptive Statistics for Two Quantitative Variables



Analyzing scatterplots is an important skill. In this graph, we see that the points seem to follow the linear pattern reasonably well and are reasonably close to the line. Shorter men on the left tend to have lower weights than taller men on the right. The line goes up from left to right. We call this a “positive linear relationship”, or a “positive correlation”. If the line goes down from left to right, we would call that a “negative linear relationship”, or a “negative correlation”.



Summary Statistics [Switch Variables](#)

Statistic	Men Ht (in)	Men Wt (Lbs)
Mean	68.335	172.550
Standard Deviation	3.020	26.327
Sample Size	40	
Correlation	0.522	
Slope	4.553	
Intercept	-138.607	

Scatterplot Controls

☒ Show Regression Line

We see that StatKey has given us the mean and standard deviation of each data set (heights and weights). It has also given us the sample size (n) of 40. There were 40 ordered pairs (40 heights and 40 weights from the same 40 men). The number next to the word "Correlation" is 0.522. This is called the "correlation coefficient" (r) and is an important statistic in measuring the direction and strength of the linear relationship. Here are some general guidelines for understanding the correlation coefficient " r ".

Correlation Coefficient (r)

The correlation coefficient (r) is a number between -1 and +1 that measures the strength and direction of correlation. The correlation coefficient is an extremely difficult calculation that is very time consuming. Like most statistics, it is better to use a computer program like StatKey or Statcato to calculate it.

If the r is negative, the regression line will go down from left to right. If you remember from algebra classes, this means the line has a negative slope. If the r is positive, the regression line will go up from left to right. This means the line has a positive slope. The closer r is to +1 or -1, the stronger the relationship. This means the points are very close to the line. The closer r is to zero, the weaker the relationship. The points are very far from the line. It is important to always look at the scatterplot with the r -value. Do not just look at an r -value without looking at the scatterplot. These are not strict rules, but general guidelines. A scatterplot with many points and a 0.7 r -value can mean something different from a scatterplot with only a few points and a 0.7 r -value.

- If r is close to +1 (like $r = +0.893$) → Strong, Positive Correlation (line going up from left to right (positive slope) and the points in scatterplot are close to line) ,
($r \approx +0.6, +0.7, +0.8, +0.9$ usually indicate pretty strong positive correlation)
- If r is close to -1 (like $r = -0.916$) → Strong Negative Correlation (line going down from left to right (negative slope) and the points in the scatterplot are close to the line)
($r \approx -0.6, -0.7, -0.8, -0.9$ usually indicate pretty strong negative correlation)
- If r close to zero (like $+0.037$ or -0.009) → No linear correlation. Points in the scatterplot do not follow any linear pattern. There still could be a nonlinear curved pattern though.
($r \approx \pm 0.1, \pm 0.0$ usually indicate no linear correlation)
- If $r \approx \pm 0.2, \pm 0.3$ usually indicate very weak linear correlation. There is some linear pattern but the points are very far from the regression line.
- If $r \approx \pm 0.4, \pm 0.5$ usually indicate moderate linear correlation. There is a linear pattern and points are only moderately close to the regression line.



In the men's height and weight example, the r -value was +0.522. This tells us that there is a moderate positive linear relationship (or moderate positive correlation) between the height and weight of these men.

Important Note: Remember relationships or associations do not imply causation. Just because there is a positive linear relationship between the height and weight of these men, it does not give me the right to say that the height causes a man to have a certain weight. There are many confounding variables involved.

Correlation \neq Causation

Coefficient of Determination (r^2)

If you square the r -value, you get the coefficient of determination. This statistic tells us the percentage of variability in the response variable (Y) that can be explained by the explanatory variable (X). In general, the higher the r^2 percentage, the stronger the relationship.

StatKey does not calculate r^2 for us, but it is not a difficult calculation. If we square the r -value, we get the following.

$$r^2 = (0.522)^2 = 0.522 \times 0.522 \approx 0.272 \text{ or } 27.2\%$$

So about 27.2% of the variability in the men's weights can be explained by the relationship with their heights.

Slope

The slope of the regression line is an important statistic in correlation and regression. It is a difficult calculation. If you are wondering how it is calculated, here is the formula the computer used.

$$\text{Slope of the Regression Line} = \frac{\text{Correlation Coefficient} \times \text{Standard Deviation of } Y}{\text{Standard Deviation of } X}$$

The slope is the amount of increase or decrease in Y for every 1-unit increase in X (per unit of X). If the slope is negative, then it is a "decrease" in Y and if the slope is positive, it is an "increase" in Y .

In this problem, StatKey gave us the slope as 4.553. Notice this is a positive slope so is indicating an increase in Y . The slope tells us that the weights of the men in the data set are increasing 4.553 pounds on average for every 1 inch taller they get. Another way to say that is that the weights are increasing on average 4.552 pounds per inch.

Y-intercept

The Y-intercept is another difficult calculation. In case you are wondering, here is the formula the computer used to calculate the Y-intercept. You must calculate the slope first, before you can find the Y-intercept.

$$\text{Y-intercept of Regression Line} = \text{Mean of } Y \text{ values} - (\text{Slope} \times \text{Mean of } X \text{ values})$$

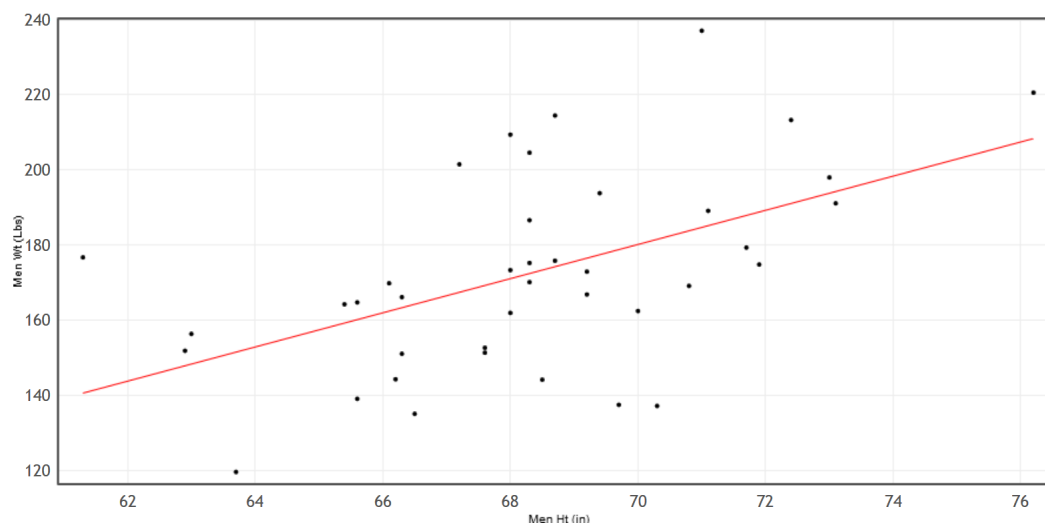
Y-intercepts can be difficult because they do not always make sense in context. The definition of a Y-intercept is the predicted Y -value when X is zero. StatKey calculated the Y-intercept for the height and weight data as -138.607. So by definition, the predicted average height of men that are zero inches tall is negative 138.607 pounds. That does not make sense.

In many situations (like heights of men), it is impossible for the X to be zero. Look at the scatterplot again for the height and weight data.



StatKey Descriptive Statistics for Two Quantitative Variables

Custom Dataset ▾ Show Data Table Edit Data Upload File Change Column(s)



Notice that the points in the scatterplot have X values between about 61 inches and about 76 inches. This is called the scope of the X-values. The accuracy of this regression line is based on X values between about 61 and 76 inches. If we use this data to predict a man's weight from his height, we should only use heights in the scope (between 61 and 76). Going outside the scope is called extrapolation and can result in bad errors. So let us get back to the Y-intercept. The Y intercept is plugging in zero for X. Notice zero is not in the scope of the X values, so is an extrapolation. That means we will not expect the Y-intercept to make sense in this context. The number is correct and is important for the regression line accuracy, but a man cannot have a height of zero.

Some Y-intercepts do make sense in context. Suppose we are looking at the number of months a company has been in business (X) and their monthly revenue in thousands of dollars (Y). The Y-intercept may represent their starting capital at month zero or the amount of money the company had when they started their business.

Regression Line and Predictions

The regression line is also called the "line of best fit" or the "line of least squares". It minimizes the vertical distances between the points in the scatterplot and regression line itself. If there is correlation between the variables, then the regression line is also a prediction formula. If you plug in an X value into the equation for X, you can solve for Y and get a predicted Y value. The regression line is represented by the following formula.

$$\hat{Y} = (\text{Y-intercept}) + (\text{Slope}) X$$

Plugging in our Y intercept (-138.607) and our slope (4.553), we get the following equation.

Regression Line for Heights and Weights of men in the health data: $\hat{Y} = -138.607 + 4.553 X$

The \hat{Y} refers to the "predicted Y value" which can be very different from the actual Y values in the data set. You may also see computer programs put in the variable names for X and \hat{Y} .

Weights in pounds = $-138.607 + 4.553$ (Heights in inches)

We said already that there was a moderate correlation between the heights and weights of these men. So we should be able to use the formula to make a prediction.

Use the regression line equation to predict the average weights of men that are 73 inches tall. Remember Y represents weight and X represents height. Simply plug in 73 for X and solve for Y. Remember to follow the order of operations. Multiply the X value by the slope first, before you add it to the Y-intercept. Also, be aware of negative Y-intercepts and negative slopes.



$$\hat{Y} = -138.607 + 4.553 X$$

$$\hat{Y} = -138.607 + 4.553 (73)$$

$$\hat{Y} \approx -138.607 + 332.369$$

$$\hat{Y} \approx +193.762$$

Therefore, we predict that the average weight of men that are 73 inches tall is about 193.8 pounds. Be careful of applying this prediction to all men. This data came from sample data and may not reflect the heights of all men on earth.

Calculating Correlation Graphs and Statistics with Statcato

We can also make scatterplots and calculate correlation statistics with Statcato. Copy and paste the men's height and weight data into two columns of Statcato. Go to the "statistics" menu, click on "correlation and regression" and then click on "linear". Click on the height to be the X-variable and the weight to be the Y-variable and then push "add series". Check the box that says "show scatterplot" and the box that says "show regression line". Statcato also has the capability of making residual plots. These are more advanced kinds of graphs that are studied in regression analysis. Check the box that says, "Show residual plots", the box that says "residuals vs x-variable", and the box that says "histogram of the residuals". Now push "OK".

Linear Correlation and Regression

Help F1

Inputs

Independent/dependent variable series

Select the independent (x) and dependent (y) variables of a regression

X variable: C1 Men Ht (in)

y variable: C2 Men Wt (Lbs)

Add Series

Select the series to be removed: Remove

Clear Input List

Significance

Significance level: 0.05

☒ Show a scatterplot for all pairs of data values

Scatterplot Options

X-axis Label: x

Y-axis Label: y

Plot Title: Scatterplot

☒ Show legend

☒ Show regression line

☒ Show Residual Plots

Residual Plot Options

☒ Residuals vs. X Variable

☐ Residuals vs. Predicted (Fitted) Values

☐ Normal Probability Plot of Residuals

☒ Histogram of Residuals

☐ Residuals vs. Observation Order

OK Cancel



Correlation and Regression: Significance level = 0.05

Series: C1 Men Ht (in), C2 Men Wt (Lbs)

$x = \text{C1 Men Ht (in)}$

$y = \text{C2 Men Wt (Lbs)}$

Sample size $n = 40$

Degrees of freedom = 38

Correlation:

$H_0: \rho = 0$ (no linear correlation)

$H_1: \rho \neq 0$ (linear correlation)

	Test Statistic	Critical Value
r	0.5222	± 0.3120
t	3.7750	± 2.0244

p-Value = 0.0005

Regression:

Regression equation $Y = b_0 + b_1x$

$b_0 = -138.6070$

$b_1 = 4.5534$

Variation:

Explained variation = 7372.6464

Unexplained variation = 19659.0136

Total variation = 27031.66

Coefficient of determination $r^2 = 0.2727$

Standard error of estimate = 22.7452

Some of the information in this printout refers to the correlation hypothesis test that we will study in chapter five. Notice Statcato gave us the correlation coefficient r of 0.522 and the coefficient of determination $r^2 = 0.2727$ (27.27%). The slope is given as $b_1 = 4.5534$ and the Y-intercept is given as $b_0 = -138.6070$. Notice these are the same numbers as StatKey.

There is one statistic on the Statcato printout that was not on the StatKey printout that is important.

Standard error of estimate = 22.7452

This statistic is called the standard deviation of the residual errors (s_e). It measures the average vertical distance that points in the scatterplot are from the regression line. It also tells us the average prediction error for predictions made in the scope of the X-values. The units of the standard deviation of the residual errors is the same as the Y-variable (pounds). This statistic tells us the following.

The points in the scatterplot are 22.7452 pounds on average from the regression line.

If we use the regression line and the height of a man to predict the weight, our prediction could have an average error of 22.7452 pounds.



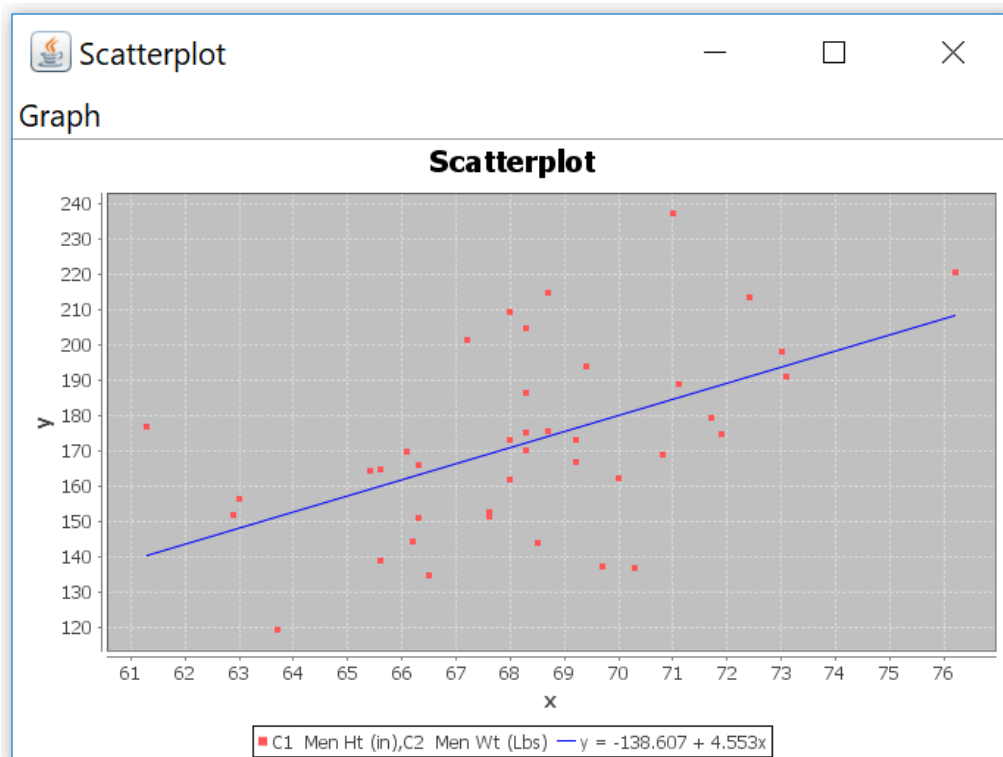
Remember the prediction we made earlier. We predicted that the average weight of men that are 73 inches tall is about 193.8 pounds. Well that prediction could be off by 22.7452 pounds on average.

A “residual” is the vertical distance that each point is from the regression line. Suppose a point has an ordered pair (X , Y). The point on the regression line with the same X value would have an ordered pair (X , \hat{Y}). To calculate a residual the computer subtracts the predicted \hat{Y} value from the actual Y value of the point in the scatterplot. This gives the vertical distance that point is from the regression line.

$$\text{Residual} = Y - \hat{Y}$$

The standard deviation of the residual errors is an average of the residuals. The actual formula is shown below. Notice that we divide by $n - 2$ instead of $n - 1$ because there were two data sets. This again is called the degrees of freedom and will be discussed more in later chapters.

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$



Notice Statcato also gave us a scatterplot of the data with the regression line drawn. The regression line formula is at the bottom of the graph.

Practice Problems Section 4G

1. How can tell which variable should be the explanatory variable and which variable should be the response variable?
2. How can we use the correlation coefficient (r) to determine if there is strong positive correlation? How can we use the correlation coefficient (r) to determine if there is strong negative correlation? How can we use the correlation coefficient (r) to determine if there is no correlation?
3. What is the definition of the coefficient of determination (r^2)?



4. What are the two definitions for the standard deviation of the residual errors (s_e)?
5. What is the definition of the slope of the regression line?
6. What is the definition of the y-intercept of the regression line?
7. What is extrapolation? Why should we avoid extrapolation?

(#8-16) Directions: Go to www.matt-teachout.org, click on "Statistics" and then "Data Sets". Open the indicated data. Copy and paste the two indicated columns of quantitative data next to each other on a new Excel spreadsheet. Then copy the two columns together. Now go to www.lock5stat.com and click on StatKey. Under the "Descriptive Statistics and Graphs" menu click on "Two Quantitative Variables". Click on "Edit Data" and paste the two columns together into StatKey. Then answer indicated questions.

8. Open the cigarette data. Let the explanatory variable (X) represent the amount of nicotine (milligrams) and the response variable (Y) represent the amount of tar (milligrams).
 - a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
 - b) Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
 - c) Find the slope of the regression line. Write a sentence to explain the slope.
 - d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
 - e) The standard deviation of the residual errors was 1.3 mg. Explain the two meanings of this statistic.
 - f) Use the regression line formula to predict the amount of tar if a cigarette contains 1.2 mg of nicotine. How much error could there be in this prediction.
9. Open the cigarette data. Let the explanatory variable (X) represent the amount of nicotine (mg) and the response variable (Y) represent the amount of carbon monoxide in parts per million (PPM).
 - a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
 - b) Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
 - c) Find the slope of the regression line. Write a sentence to explain the slope.
 - d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
 - e) The standard deviation of the residual errors was 2.3 PPM. Explain the two meanings of this statistic.
 - f) Use the regression line formula to predict the amount of carbon monoxide if a cigarette contains 1.2 mg of nicotine. How much error could there be in this prediction.
10. Open the health data. Let the explanatory variable (X) represent the systolic blood pressure (mm of Hg) and the response variable (Y) represent the diastolic blood pressure (mm of Hg). Use the combined columns with 80 randomly selected adults. Do not separate by gender.
 - a) Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
 - b) Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
 - c) Find the slope of the regression line. Write a sentence to explain the slope.
 - d) Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
 - e) The standard deviation of the residual errors was 7.4579 mm of Hg. Explain the two meanings of this statistic.
 - f) Use the regression line formula to predict the diastolic blood pressure of a person who has a systolic blood pressure of 130. How much error might there be in that prediction?



11. Open the health data. Let the explanatory variable (X) represent the waist size in centimeters and the response variable (Y) represent the weight in pounds. Use the combined columns with 80 randomly selected adults. Do not separate by gender.

- Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
- Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
- Find the slope of the regression line. Write a sentence to explain the slope.
- Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
- The standard deviation of the residual errors was 14.6809 pounds. Explain the two meanings of this statistic.
- Use the regression line formula to predict the weight of a person who has a waist size of 100 cm. How much error might there be in that prediction?

12. Open the health data. Let the explanatory variable (X) represent the age in years and the response variable (Y) represent the cholesterol in milligrams per deciliter (mg/dL). Use the combined columns with 80 randomly selected adults. Do not separate by gender.

- Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
- Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
- Find the slope of the regression line. Write a sentence to explain the slope.
- Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
- The standard deviation of the residual errors was 255.3625 mg/dL. Explain the two meanings of this statistic.
- Use the regression line formula to predict the cholesterol of a person that is 40 years old. How much error might there be in that prediction?

13. Open the bear data. Let the explanatory variable represent the age of the bear in months and the response variable represent the length of the bear in inches.

- Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
- Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
- Find the slope of the regression line. Write a sentence to explain the slope.
- Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
- The standard deviation of the residual errors was 7.51 inches. Explain the two meanings of this statistic.
- Use the regression line formula to predict the length of a bear that is 24 months old. How much error might there be in that prediction?

14. Open the bear data. Let the explanatory variable represent the neck circumference of the bear and the response variable represent the weight of the bear in pounds.

- Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
- Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
- Find the slope of the regression line. Write a sentence to explain the slope.
- Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
- The standard deviation of the residual errors was 43.9 pounds. Explain the two meanings of this statistic.



- f) Use the regression line formula to predict the weight of a bear that has a neck circumference of 24 inches. How much error might there be in that prediction?
15. Open the car data. Let the explanatory variable (X) represent the weight of the car in tons and the response variable (Y) represent the gas mileage in miles per gallon.
- Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
 - Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
 - Find the slope of the regression line. Write a sentence to explain the slope.
 - Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
 - The standard deviation of the residual errors was 2.8516 mpg. Explain the two meanings of this statistic.
 - Use the regression line formula to predict the mpg for a car that weighs 3 tons. How much error might there be in that prediction?
16. Open the car data. Displacement is the amount of liquid in cubic centimeters forced out by the piston. Let the explanatory variable (X) represent the horsepower of the car and the response variable (Y) represent the displacement of the car (cc).
- Look at the scatterplot and the correlation coefficient (r). Describe the strength and direction of the linear relationship.
 - Square the correlation coefficient in StatKey to calculate r^2 . This is also called the coefficient of determination. Write a sentence to explain r^2 .
 - Find the slope of the regression line. Write a sentence to explain the slope.
 - Find the y-intercept. Write a sentence to explain the y-intercept. Does the y-intercept make sense in the context of this data?
 - The standard deviation of the residual errors was 44.138 cubic centimeters. Explain the two meanings of this statistic.
 - Use the regression line formula to predict the number of cc's of displacement for a car with 120 horsepower. How much error might there be in that prediction?
-

Section 4H – Quantitative Relationships: The Correlation Test

We saw in the last section, that two quantitative samples are related if their correlation coefficient (r) is close to 1 or -1 . When the correlation coefficient (r) is close to zero, the two quantitative samples are not related. How does this apply to populations? What if we want to determine if there is a relationship between two quantitative variables in a population? For this, we will need to look at the correlation hypothesis test.

The Correlation Hypothesis Test

If the sample correlation coefficient (r) is zero, tells us that the two quantitative samples are not related. For populations, we need to look at the population correlation coefficient “rho” (ρ). While this looks like a “p”, it is not. It is the Greek letter “rho” and represents the population correlation coefficient.

If you recall from the last section, the correlation coefficient is related to the slope of the regression line.

$$\text{Sample Slope } b_1 = \frac{(\text{correlation coefficient times standard deviation of the } y \text{ values})}{\text{standard deviation of the } x \text{ values}} = \frac{(r \times S_y)}{S_x}$$

So if there is no correlation between variables the correlation coefficient and the slope both go to zero. This principle applies to populations as well. As the population correlation coefficient “rho” (ρ) goes to zero, the population slope “Beta 1” (β_1) also goes to zero.



Correlation Test Null and Alternative Hypothesis

There are several ways of writing the null and alternative hypothesis for a correlation hypothesis test. We can use the population correlation coefficient “rho” (ρ) or the population slope “Beta 1” (β_1). We can also specify positive or negative correlation. Remember the correlation coefficient and the slope always have the same sign.

To show positive correlation the correlation coefficient should be close to +1 (greater than zero) and the slope should also be significantly positive (greater than zero). A positive relationship is also called a “direct” relationship. As the X variable increases, the Y variable also tends to increase. As the X variable decreases, the Y variable also tends to decrease.

To show negative correlation the correlation coefficient should be close to -1 (less than zero) and the slope should also be significantly negative (less than zero). A negative relationship is also called an “indirect” or “inverse” relationship. As the X variable increases, the Y variable also tends to decrease. As the X variable decreases, the Y variable also tends to increase.

Note about Statcato: Statcato only has the option for the two-tailed correlation test and cannot specify positive correlation (right-tailed) or negative correlation (left-tailed) hypothesis tests.

Two-Tailed Correlation Test: For determining if variables are related or not. Does not specify if the direction is positive or negative.

$H_0 : \rho = 0$ (The two quantitative variables in the population are not related.)

$H_A : \rho \neq 0$ (The two quantitative variables in the population are related.)

OR

$H_0 : \beta_1 = 0$ (The two quantitative variables in the population are not related.)

$H_A : \beta_1 \neq 0$ (The two quantitative variables in the population are related.)

Right-Tailed Correlation Test: For determining if variables have a positive (or direct) relationship or not. Notice the alternative hypothesis symbol “>” points to the right.

$H_0 : \rho = 0$ (The two quantitative variables in the population are not related.)

$H_A : \rho > 0$ (The two quantitative variables in the population have a positive (direct) relationship.)

OR

$H_0 : \beta_1 = 0$ (The two quantitative variables in the population are not related.)

$H_A : \beta_1 > 0$ (The two quantitative variables in the population have a positive (direct) relationship.)

Left-Tailed Correlation Test: For determining if variables have a negative (or inverse) relationship or not. Notice the alternative hypothesis symbol “<” points to the left.

$H_0 : \rho = 0$ (The two quantitative variables in the population are not related.)

$H_A : \rho < 0$ (The two quantitative variables in the population have a negative (inverse) relationship.)

OR

$H_0 : \beta_1 = 0$ (The two quantitative variables in the population are not related.)

$H_A : \beta_1 < 0$ (The two quantitative variables in the population have a negative (inverse) relationship.)

T-test statistic

The relationship between correlation and the slope of the regression line is highlighted in the test statistic. For a correlation test, you can use either the correlation coefficient “r” or a T-test statistic. I prefer the T-test statistic. The null hypothesis is that there is not a relationship between the quantitative variables. This would indicate that the correlation coefficient and the slope would be close to zero. So the T-test statistic counts how many standard error



the slope is from zero. If the T-test statistic is positive, the slope will be above zero and if the T-test statistic is negative, the slope will be below zero.

$$\text{T-test statistics for correlation} = \frac{(\text{slope}-0)}{\text{standard error}}$$

T-test statistics sentence for Correlation: The number of standard errors that the slope of the regression line is above or below zero.

As with all test statistics, we will want to see if the T-test statistic falls in a tail determined by the critical value or values. If so, the sample data significantly disagrees with the null hypothesis and the slope is significantly different from zero. If the T-test statistic does not fall in a tail determined by the critical value or values, then the sample data does not significantly disagree with the null hypothesis and the slope is not significantly different from zero.

Residual Errors

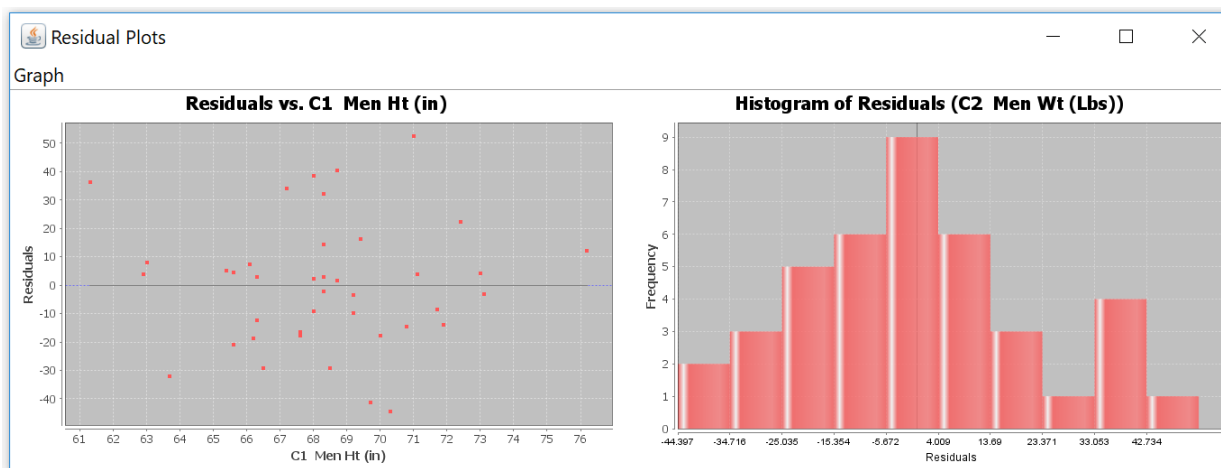
The correlation test has many assumptions. Some of the assumptions are centered on the understanding of “residuals” or “residual errors”. We learned in the last section that a residual is the vertical distance between each point in the scatterplot and the regression line. To calculate a residual, the computer subtracts the actual y coordinate of the point minus the predicted \hat{y} value on the regression line. We also saw that the average of all the residuals is called the standard deviation of the residual errors (S_e). This tells us the average vertical distance that the data is from the regression line and the average prediction error.

$$\text{Residual} = y - \hat{y}$$

$$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$$

Besides the standard deviation of the residual errors, there are also residual graphs that statisticians often like to examine when doing a correlation test. We will only look at two. They are the “histogram of the residuals errors” and the “residual plot verses the x-values”.

Here is an example. These graphs were created with Statcato. The explanatory variable (X) was the height of men and the response variable (Y) was the weight of men.



Residual Plot

The graph on the left is called the “residual plot verses the x-variable”. This graph shows the vertical distances that each point is from the regression line. A point that is 40 above the line will have a residual of +40. A point that is 19 below the line will have a residual of -19. The zero line represents the regression line since points on the regression



line have a residual of zero. We want the residual plot to be evenly spread out. When the points are evenly spread out, our standard deviation of the residuals is a consistent measure of spread. When the residual plot is not evenly spread out, you will see parts of the x-axis where all the points are very close and other parts of the x-axis where the points are very far away. This is an uneven spread (or fan shaped). We want the standard deviation to be a consistent measure of spread for all x value in the scope. If the points are close for some x values, then the standard deviation will be an overestimate of the variability for those x values. Similarly, if the points are far away for other x values, then the standard deviation will be an underestimate of the variability for those x values. Residual plots can be very difficult to read. I tell my intro students to put all the points on the left side of the graph between your fingers. Now put all the points on the right side of the graph between your fingers. If your fingers are about the same width on both the left and right side, you are probably ok. The data is evenly spread out and the standard deviation is a consistent measure of spread (variability). If your fingers are much closer on one side than the other, that may indicate a fan shape or uneven spread. In that case, the standard deviation is not a consistent measure of variability. Notice that points with an x-value greater than 72 are much closer to the regression line than those below 72. This could indicate an uneven spread (fan shape). This also could indicate that the regression line predictions are more accurate for taller men in the data (over 72 inches) and less accurate for shorter men in the data.

Histogram of the Residuals

The graph on the right is called the “histogram of the residuals”. Remember that the calculation of the regression line uses the mean and standard deviation. If you remember from previous chapters, the mean and standard deviations are only accurate for normal data. We could check the shape of each data set separately, but instead we prefer to check the shape of the residuals. The histogram of the residuals should be normal (bell shaped). It should also be centered close to zero. Statcato gives a dark vertical line at zero for this purpose. This line should be close to the highest bar in the histogram. This histogram above passes both criteria.

Let us look at the assumptions for a correlation test.

Correlation Test Assumptions

1. The quantitative ordered pair data should be collected randomly or be representative of the population. *(The two samples usually have different units, but must have a one-to-one pairing).*
2. Data values within the sample should be independent of each other. *(The two samples are not independent since they are ordered pair. The individual data values within each sample should be independent. If you have small simple random sample from a large population, then the data values are probably not related.)*
3. The sample size should be at least 30. *(There should be 30 or more ordered pairs.)*
4. The scatterplot and correlation coefficient (r) should show some linear pattern. *(The correlation coefficient (r) should not be close to zero.)*
5. There should be no influential outliers in the scatterplot. *(If your correlation coefficient is close to 1 or -1 , then you probably have no influential outliers. Remember to look for outliers on the scatterplot. A residual plot magnifies the distances, so everything looks like an outlier in a residual plot.)*
6. The histogram of the residuals should be nearly normal.
7. The histogram of the residuals should be centered close to zero. *(The zero line should be touching the highest bar in the histogram, or at least very close to the highest bar.)*
8. The residual plot verses the x variables should be evenly spread out with no fan shape or sideways “V” pattern. *(Put all the points in the residual plot between your fingers on the left side of the graph. Now put all the points in between your fingers on the right side. If your fingers are about the same width apart on the left and right side, the graph is close to evenly spread out.)*



Correlation Test Example 1

Let us use Statcato and the random “Health” data at www.matt-teachout.org to test the claim that there is no relationship between the age of a man and his cholesterol. In the Health data, we have the ages and cholesterol of forty randomly selected men. We will designate the age to be the explanatory variable (X) and the cholesterol to be the response variable (Y). Let us use a 5% significance level.

We can write the null and alternative hypothesis in one of two ways. We can use the population correlation coefficient “rho” (ρ) or the population slope “beta 1” (β_1). Remember our claim is “not related” so that must be the null hypothesis. Since positive or negative relationship was not mentioned, we will assume this is the general two-tailed test.

$H_0 : \rho = 0$ (The age and cholesterol of men are not related.) CLAIM

$H_0 : \rho \neq 0$ (The age and cholesterol of men are related.)

OR

$H_0 : \beta_1 = 0$ (The age and cholesterol of men are not related.) CLAIM

$H_0 : \beta_1 \neq 0$ (The age and cholesterol of men are related.)

Copy and paste the men’s age and cholesterol data into two columns of Statcato. Go to the “statistics” menu, click on “correlation and regression” and then click on “linear”. Click on the men’s age to be the X-variable and the men’s cholesterol to be the Y-variable and then push “add series”. Check the box that says “show scatterplot” and the box that says “show regression line”. Statcato also has the capability of making residual plots. Check the box that says, “Show residual plots”, the box that says “residuals vs x-variable”, and the box that says “histogram of the residuals”. Now push “OK”. Here is the Statcato printout, with the test statistic, P-value, correlation coefficient and all of the graphs.

Note: Some versions of Statcato do not have residual plots.



Correlation and Regression: Significance level = 0.05

Series: C15 Men Age (years),C22 Men Chol

x = C15 Men Age (years)

y = C22 Men Chol

Sample size $n = 40$

Degrees of freedom = 38

Correlation:

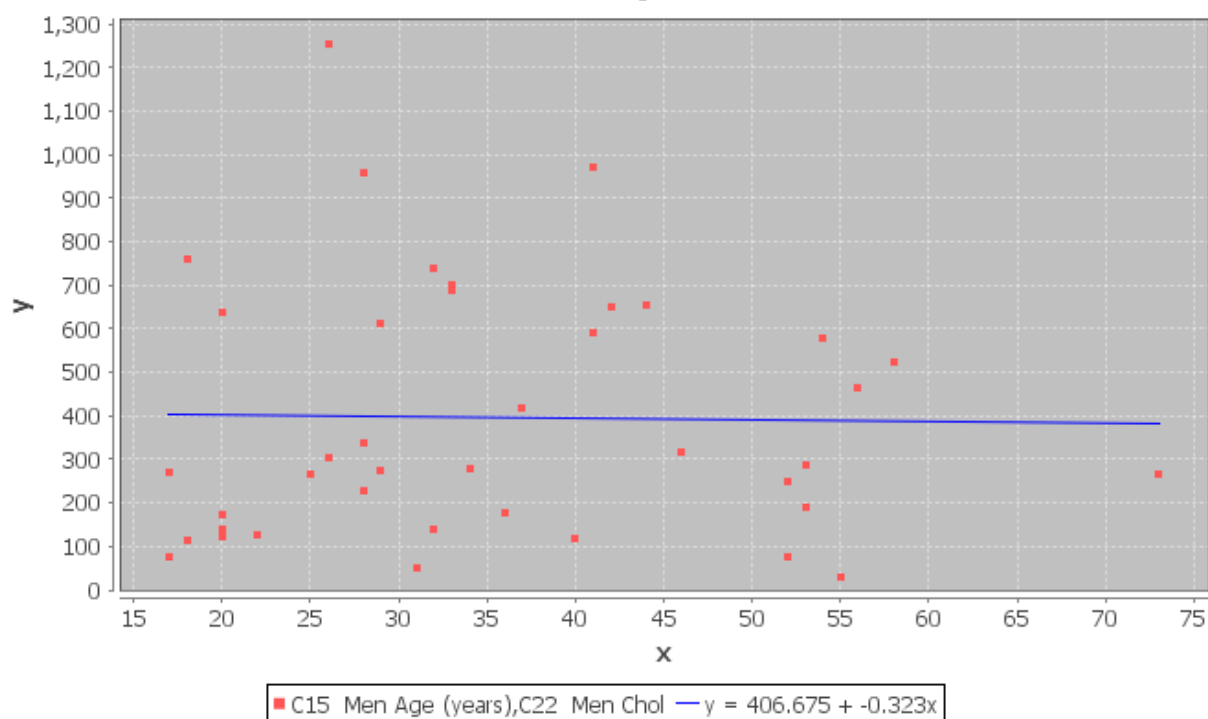
$H_0: \rho = 0$ (no linear correlation)

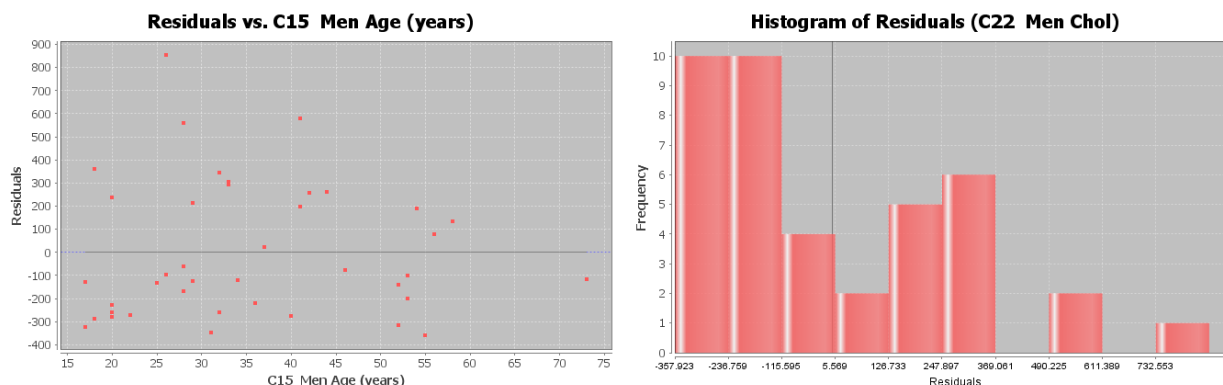
$H_1: \rho \neq 0$ (linear correlation)

	Test Statistic	Critical Value
r	-0.0154	± 0.3120
t	-0.0948	± 2.0244

p-Value = 0.9250

Scatterplot





Let us start by checking the assumptions for the men's age and cholesterol problem. Notice that this data fails many of the assumptions. That means our hypothesis test is compromised. We should also not use this regression line to make predictions about men's cholesterol.

1. Two quantitative ordered pair random samples. **Yes.** Age and cholesterol are both quantitative. The data had randomly selected men with the age and cholesterol of each man. It is ordered pair data.
2. Data values within each sample should be independent of each other. **Yes.** Since there is only forty randomly selected men out of millions of men in the population, the men are not likely to be related.
3. The sample size should be at least 30. **Yes.** There was forty men in the data. This is greater than thirty.
4. The scatterplot and correlation coefficient (r) should show some linear pattern. **No.** The regression line does not seem to fit the points in the scatterplot at all and the correlation coefficient r is very close to zero.
5. There should be no influential outliers in the scatterplot. **No.** There seem to be many influential outliers in the scatterplot and the correlation coefficient r is very close to zero.
6. The histogram of the residuals should be nearly normal. **No.** The histogram of the residuals is skewed right and not normal.
7. The histogram of the residuals should be centered close to zero. **No.** The histogram of the residuals seems to be centered to the left of zero. The zero line is not touching the highest bar in the histogram.
8. The residual plot verses the x variables should be evenly spread out. **No.** The residual plot seems to show a distinct fan shape and is not evenly spread out. The points on the left side of the graph seem to have a very wide spread while the points on the right side of the graph seem to be very close.

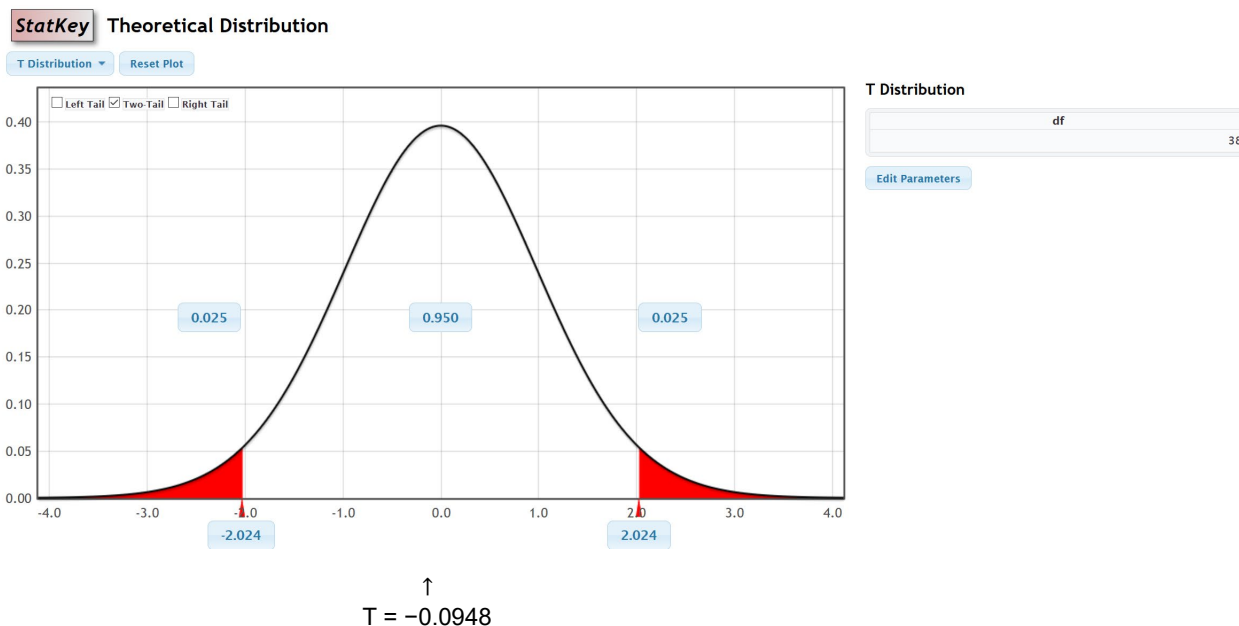
Test Statistic: $T = -0.0948$

Sentence: The slope of the regression line is 0.0948 standard errors below zero.

Our T-test statistic is -0.0948 and does not fall in either of the tails determined by the critical value. Our random sample data does not significantly disagree with the null hypothesis. This also indicates the slope is not significantly different from zero.

We put the degrees of freedom 38 into the theoretical T distribution calculator in StatKey to get the following picture.





P-value = 0.9250

Sentence: If the null hypothesis is true and there is no relationship between the age and cholesterol for men, then there is a 92.5% probability of getting the sample data or more extreme because of sampling variability.

Notice the P-value is greater than our 5% significance level. This indicates that the sample data or more extreme could have occurred because of sampling variability if the null hypothesis was true. Since sampling variability cannot be ruled out, we must fail to reject the null hypothesis.

Fail to reject the Null Hypothesis.

We have a high P-value and the null hypothesis is the claim. The sample data did not pass all of the assumptions for the correlation test.

Conclusion: There is not significant evidence to reject the claim that the age and cholesterol of men is not related.

Age and cholesterol of men are probably not related. This sample data did not provide evidence since the P-value was high and it failed many of the assumptions for the correlation test.

Example 2

We can also use randomized simulation on StatKey to determine significance and calculate the P-value. StatKey can calculate the scatterplot and the correlation coefficient and slope, but does not calculate any of the residual graphs.

We are going to be using the “mpg weight horsepower” data on www.matt-teachout.org to test the claim that there is a negative (inverse) relationship between the weight of a car and the miles per gallon of gas (mpg). We will be using a 5% significance level and assume the data met all of the assumptions.

$H_0 : \rho = 0$ (The weight and mpg of a car are not related.)

$H_A : \rho < 0$ (The weight and mpg of a car have a negative (inverse) relationship.) CLAIM

OR

$H_0 : \beta_1 = 0$ (The weight and mpg of a car are not related.)

$H_A : \beta_1 < 0$ (The weight and mpg of a car have a negative (inverse) relationship.) CLAIM



We will designate the weight of the car as the explanatory variable (X) and the miles per gallon as the response variable (Y). Copy and paste the weight of the cars and mpg into a fresh excel spreadsheet. Put the weight data on the left and the mpg on the right. Now copy both columns together.

Weight (Tons)	MPG
4.36	16.9
4.05	15.5
3.61	19.2
3.94	18.5
2.16	30
2.56	27.5
2.3	27.2
2.23	30.9

Go to www.lock5stat.com and open StatKey. Under the “Randomized Hypothesis Tests” menu, click on “Test for Slope, Correlation”. Under the “Edit Data” menu, paste in the weight and mpg columns. Since the data sets have titles, check the box that says, “Header has header row” and push OK. Under “Original Sample” we see the scatterplot, correlation coefficient (r) and the sample slope (b_1).

Edit data

Weight (Tons),MPG

4.36,16.9
4.05,15.5
3.61,19.2
3.94,18.5
2.16,30
2.56,27.5
2.3,27.2
2.23,30.9
2.83,20.3
3.14,17
2.8,21.6
3.41,16.2
3.38,20.6
3.07,20.8
3.62,18.6
3.41,18.1
3.84,17
3.73,17.6
3.96,16.5
3.83,18.2
2.56,27.5

☒ Data has header row

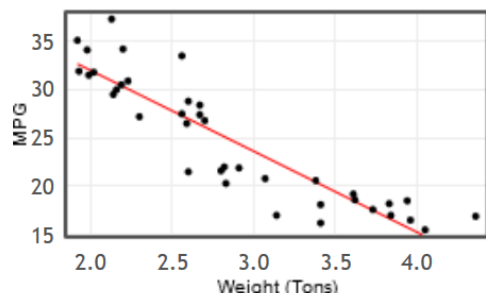
Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only two columns

Ok



Original Sample

$n = 38$, $r = -0.903$, $slope = -8.372$, $intercept = +48.74$



Let us give a quick analysis of the sample data as we did in the last section. We see that the scatterplot and the correlation coefficient (r) show a strong negative relationship between the samples. Notice that $r = -0.903$ and is close to -1 . The points in the scatterplot seem to be close to the regression line and there does not appear to be any influential outliers.

The slope is -8.372 . In our last section, we saw that the slope is the amount of increase or decrease in the Y variable per unit of X. Since the slope was negative, it is a decrease. In addition, the X variable is the weight of the car in tons and the Y variable is the gas mileage in miles per gallon.

Sample Slope Sentence: For every 1 ton heavier the car, the average miles per gallon of the cars in the samples are decreasing 8.372 mpg.

Randomized Simulation

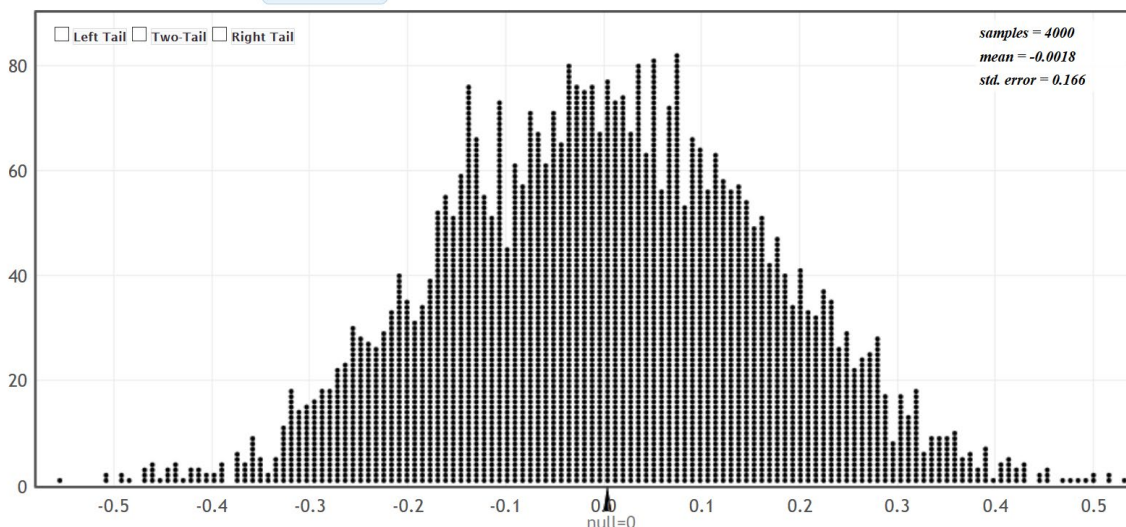
There are two ways to do the randomized simulation. We can have the computer create thousands of random samples and calculate the correlation coefficient for each. Another way is to have the computer create thousands of random samples and calculate the slope for each. At the top of the distribution, you will see we can change the setting to “correlation” or “slope”. Notice how the null hypothesis changes to reflect the setting. Click on “Generate 1000 Samples” a few times.

StatKey Randomization Test for a Slope, Correlation

Custom Dataset Show Data Table Edit Data Upload File Change Column(s)

Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Randomization Dotplot of Correlation Null hypothesis: $\rho = 0$

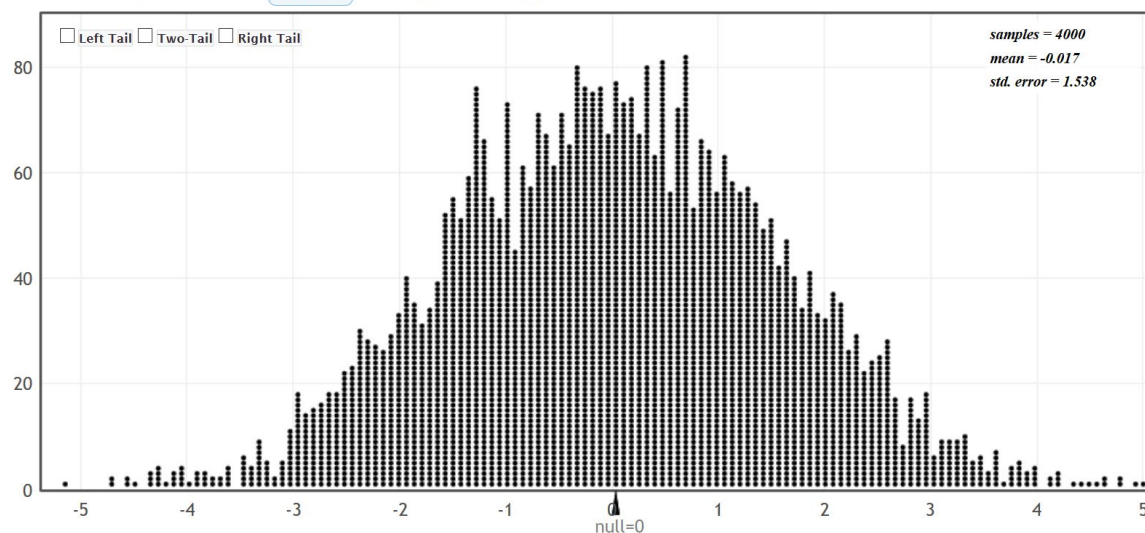


This material is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) – 10/1/18

StatKey Randomization Test for a Slope, Correlation

Custom Dataset ▾ Show Data Table Edit Data Upload File Change Column(s)
 Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

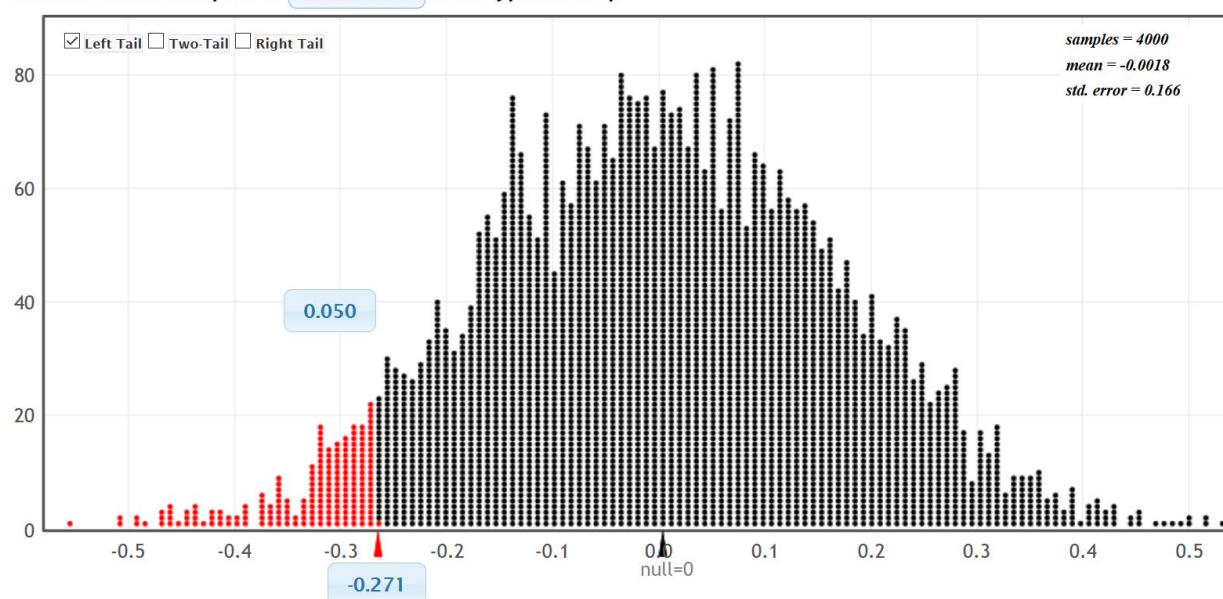
Randomization Dotplot of **Slope** ▾ Null hypothesis: $\beta_1 = 0$



Simulating with the Correlation Coefficient

Let us start with looking at the correlation coefficient simulation. These are thousands of correlation coefficients. When the setting is on "Correlation", we will need to use the "Original Sample" correlation coefficient (r) to determine significance and calculate the P-value. Since the alternative hypothesis was less than " $<$ ", this was a left-tailed test. Click on left tail. Since we are using a 5% significance level, we will put in 0.05 in the left tail proportion. Notice the simulation indicates that our "Original Sample" correlation coefficient (r) needs to be -0.271 or less to be significant. Our "Original Sample" correlation coefficient (r) is -0.903 and definitely falls in the left tail.

Randomization Dotplot of **Correlation** ▾ Null hypothesis: $\rho = 0$



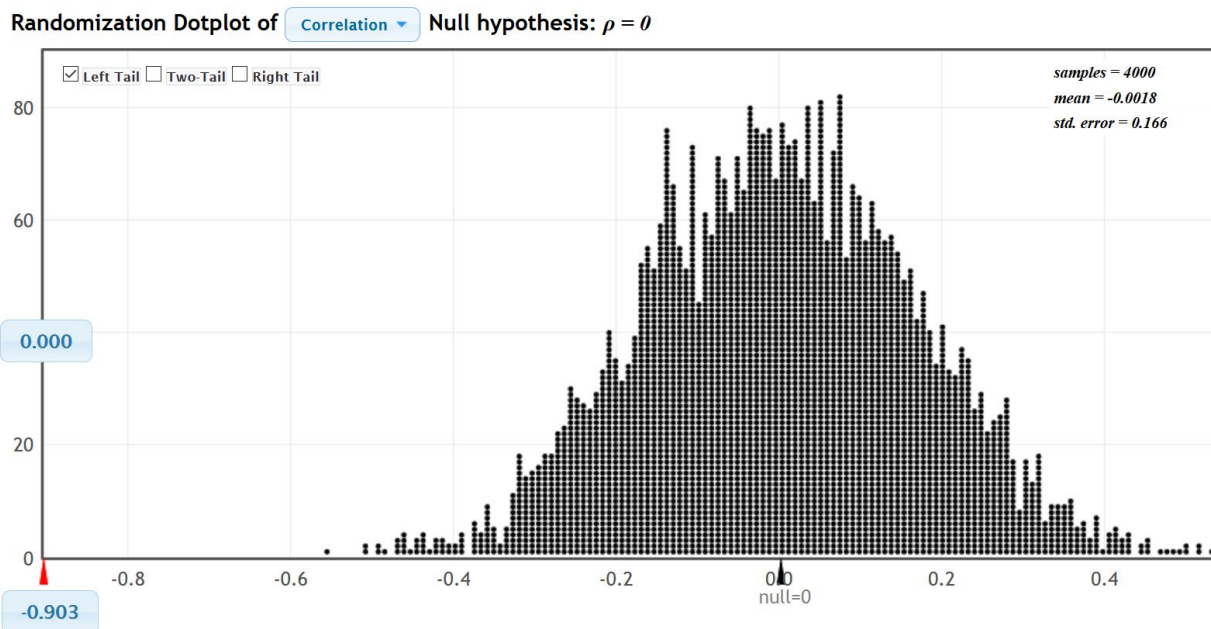
↑
 $r = -0.903$



Since our correlation coefficient r falls in the left tail of the simulation, the sample data significantly disagrees with the null hypothesis.

Calculating the P-value

The P-value is the probability of getting the sample data or more extreme by sampling variability if the null hypothesis is true. This simulated distribution is a view of sampling variability if the null is true. We just need to figure out the probability of the sample data or more extreme. Since this simulation created thousands of correlation coefficients, we will enter the real “original sample” correlation coefficient ($r = -0.903$) in the bottom box of the simulation. The left tail probability will give us the probability we are looking for. In this case, the P-value was approximately zero.

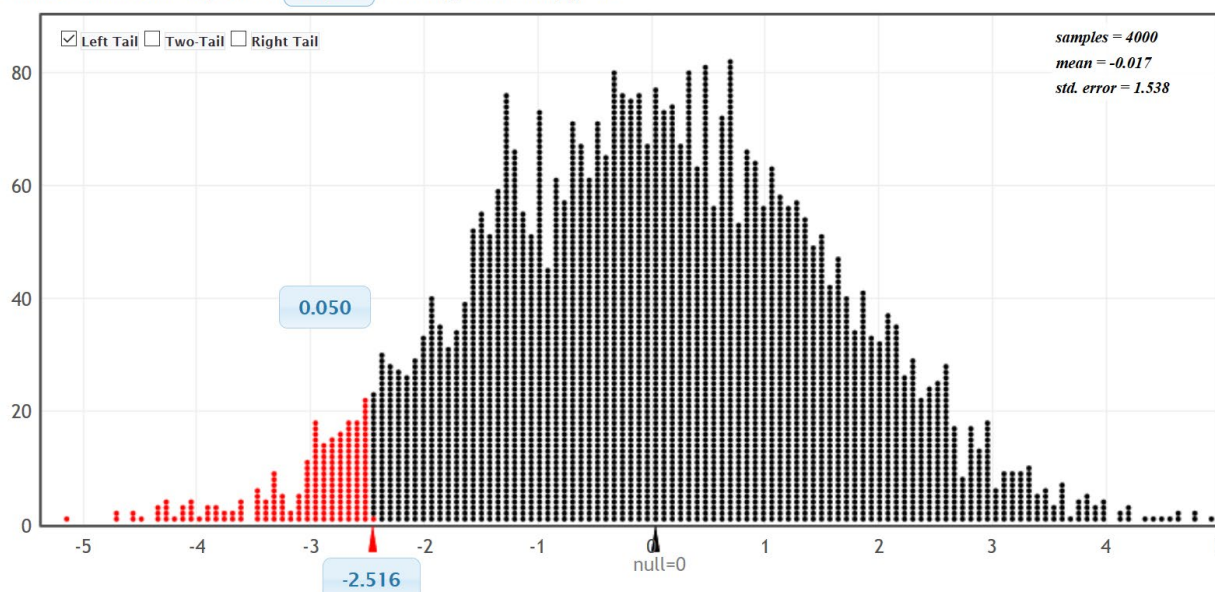


Simulating with the Slope

We can simulate with either the correlation coefficient or the slope. Here is the randomized simulation of thousands of sample slopes. Putting the 5% significance level in the tail, shows us that the real “original sample” slope needs to be -2.516 or less to be in the left tail. So if the real “original sample” slope is less than -2.516 , the sample data will significantly disagree with the null hypothesis. The real “original sample” slope is -8.372 so it does fall in the left tail.



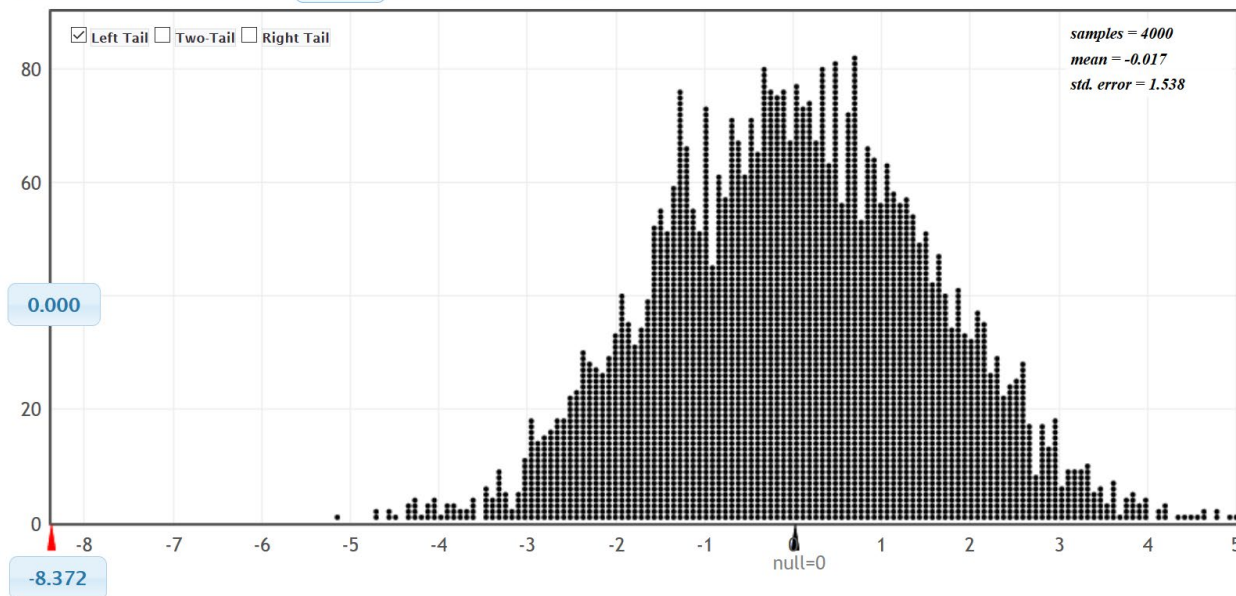
Randomization Dotplot of Slope Null hypothesis: $\beta_1 = 0$



↑
Slope = -8.372

Now let us calculate the P-value with the slope. Enter the real “original sample” slope in the bottom box in the left tail. We see that the P-value is zero. Notice this is the same P-value as we got when we simulated with the correlation coefficient.

Randomization Dotplot of Slope Null hypothesis: $\beta_1 = 0$



Notice that the original sample slope or correlation coefficient fell in the tail. So the sample data significantly disagrees with the null hypothesis. The slope is significantly different from zero.



What is the T-test statistic? Remember in a simulation, you do necessarily have to use the test statistic to judge significance. We used the sample correlation coefficient and slope to judge significance. We can calculate the T-test statistic though using the formula. Notice in the slope simulation, the approximate standard error for this simulation is 1.538. The standard error will vary between simulations though.

$$\text{T-test statistic (for the correlation test)} = \frac{(\text{Slope} - \text{Zero})}{\text{Standard Error}} = \frac{(-8.372 - 0)}{1.538} \approx -5.443$$

T-test statistics Sentence: The slope of the regression line is 5.443 standard errors below zero.

P-value ≈ 0

P-value sentence: If the null hypothesis is true and there is no relationship between the weight of a car and the miles per gallon, then there is zero probability of getting this sample data or more extreme by sampling variability.

The P-value also tells us that it is extremely unlikely for this sample data to occur because of sampling variability.

The P-value is less than our 5% significance level, so we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that there is a negative (inverse) relationship between the weight of a car and the miles per gallon. This does not imply that a heavy car causes the car to have

Notes

- Remember, if you simulate with the correlation coefficient, then you have to use the real "original sample" correlation coefficient when you calculate the approximate P-value. If you simulate with the slope, then you have to use the real "original sample" slope when you calculate the approximate P-value.
- You do not have to simulate with both the correlation coefficient and the slope. The point is that either simulation gives you approximately the same P-value.
- In all randomized simulations, there is sampling variability. Answers will vary slightly in different simulations.

Practice Problems Section 4H

(#1-10) Use either the correlation coefficient or the T-test statistic and the corresponding critical values to fill out the table.

	T-test statistic or Correlation Coefficient (r)	Sentence to explain T-test statistic or Correlation Coefficient (r)	Critical Value (T or r)	Does the T-test statistic or r-value fall in a tail determined by a critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	T = -2.441		± 1.775		
2.	r = 0.183		0.316		
3.	T = +1.166		+2.003		
4.	r = -0.799		± 0.286		
5.	T = +3.118		+2.714		
6.	r = 0.921		0.339		
7.	T = -0.852		± 2.322		
8.	r = -0.026		-0.279		
9.	T = +1.339		± 1.997		
10.	r = 0.483		+0.303		



(#11-20) Use each of the following P-values and corresponding significance levels to fill out the table.

	P-value Proportion	P- value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.521			10%			
12.	0.0426			1%			
13.	3.41×10^{-5}			5%			
14.	0.0033			1%			
15.	0.768			5%			
16.	0			10%			
17.	0.0428			5%			
18.	0.277			10%			
19.	6.04×10^{-6}			1%			
20.	0.0178			5%			

21. List the assumptions that we need to check when performing a correlation hypothesis test.

22. How can we use the scatterplot and the correlation coefficient (r) to determine if the sample data follows a linear pattern?

23. Points in the scatterplot that are far from the regression line are considered outliers, but it is difficult to know if the outliers are influential or not. How can we use the scatterplot and the correlation coefficient (r) to determine if potential outliers are influential or not?

24. Explain the two assumptions that we check by using the histogram of the residuals.

25. Explain how to determine if the residual plot is evenly spread out or not.

(#26-29) Directions: For each of the following problems, use the Statcato printouts provided to answer the following questions.

- Write the null and alternative hypothesis for the correlation test. Address the quantitative relationship and label which is the claim.
- Write a sentence to explain the strength and direction based on the correlation coefficient (r).
- Write a sentence to explain the sample slope (b_1).
- Check all of the assumptions for the correlation test. Explain your answers.
- Write a sentence to explain the T-test statistic.
- Compare the T-test statistic to the critical value. Does the test statistic fall in a tail determined by the critical value?
- Does the sample data significantly disagree with the null hypothesis? Explain your answer.
- Is the sample slope significantly different from zero? Explain your answer.
- Write a sentence explaining the P-value.
- Compare the P-value to the significance level. Could the sample data or more extreme occur by sampling variability if the null hypothesis was true or is it unlikely? Explain your answer.
- Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- Write a conclusion for the test addressing evidence and the claim.



26. Use a 5% significance level and the Statcato printout below to test the claim that there is a linear relationship between the height (X) of a man and his weight (Y). This printout came from the random health data at www.matt-teachout.org.

Correlation and Regression: Significance level = 0.05

x = C16 Men Ht (in)

y = C17 Men Wt (Lbs)

Sample size n = 40

Degrees of freedom = 38

	Test Statistic	Critical Value
r	0.5222	± 0.3120
t	3.7750	± 2.0244

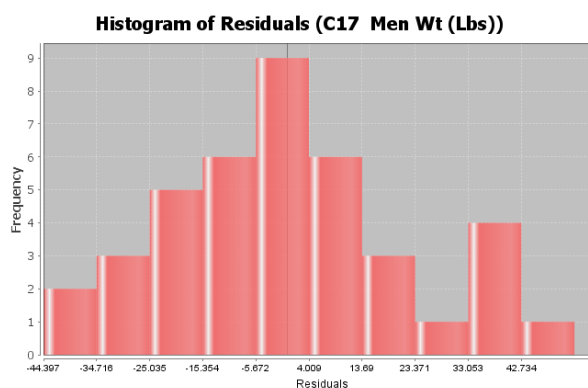
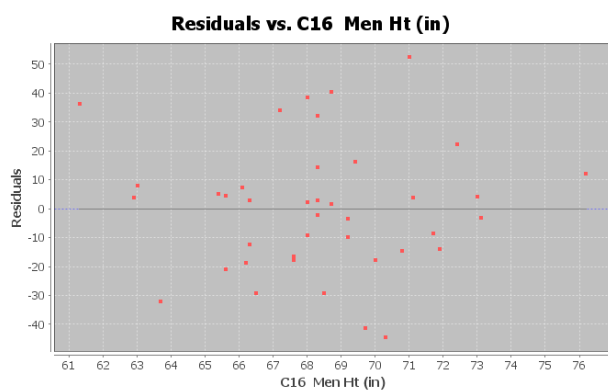
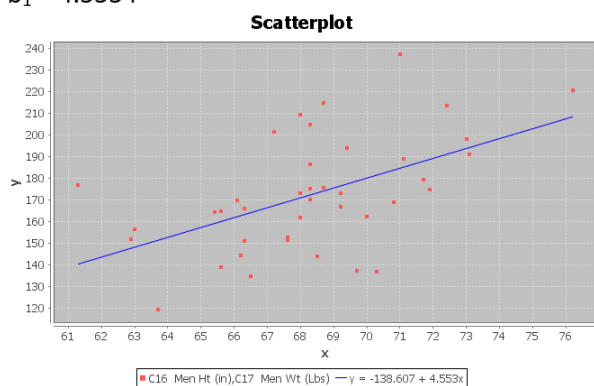
p-Value = 0.0005

Regression:

Regression equation $Y = b_0 + b_1x$

$b_0 = -138.6070$

$b_1 = 4.5534$



27. Use a 5% significance level and the Statcato printout below to test the claim that there is NO linear relationship between the systolic blood pressure (X) of a woman and her diastolic blood pressure (Y). This printout came from the random health data at www.matt-teachout.org.

Correlation and Regression: Significance level = 0.05

x = C6 Women Syst BP

y = C7 Women Diast BP

Sample size n = 40

Degrees of freedom = 38

	Test Statistic	Critical Value
r	0.7854	± 0.3120
t	7.8209	± 2.0244

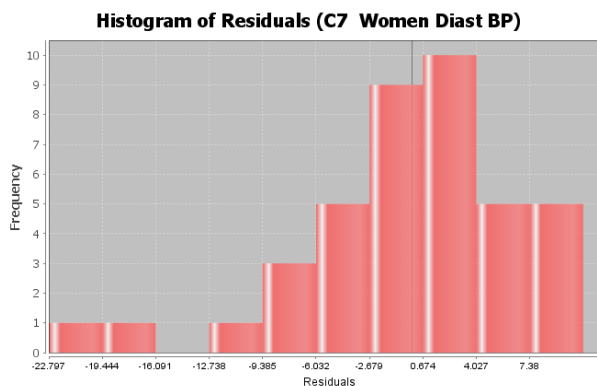
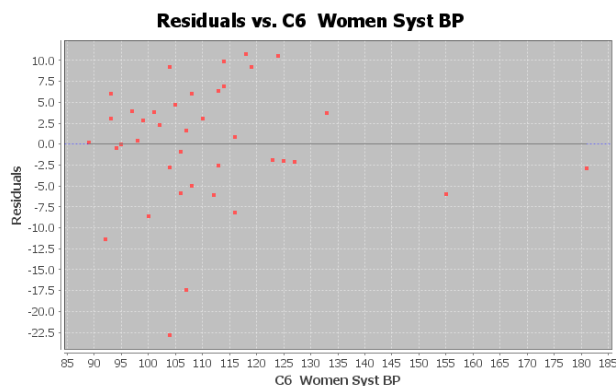
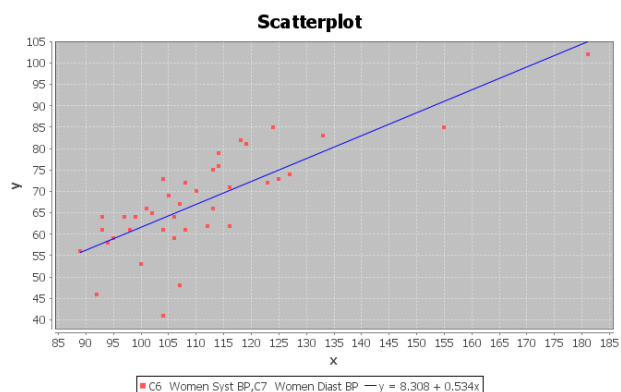
p-Value = $1.9615 \cdot 10^{-9}$

Regression:

Regression equation $Y = b_0 + b_1X$

$b_0 = 8.3079$

$b_1 = 0.5335$



28. Use a 5% significance level and the Statcato printout below to test the claim that there is a relationship between the head width (X) of a bear and its chest size (Y). This printout came from the random bear data at www.matt-teachout.org.

Correlation and Regression: Significance level = 0.05

x = C5 Head Width (In)

y = C8 Chest (in)

Sample size n = 54

Degrees of freedom = 52

	Test Statistic	Critical Value
r	0.7785	± 0.2681
t	8.9451	± 2.0067

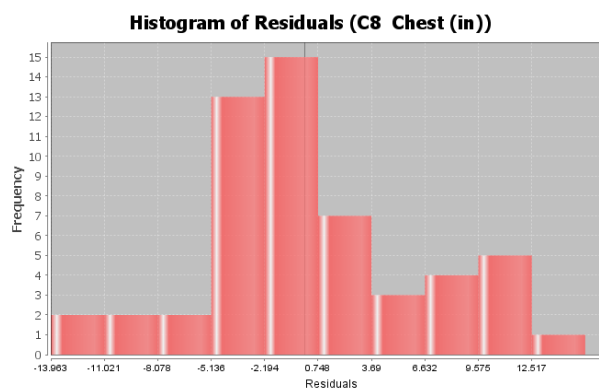
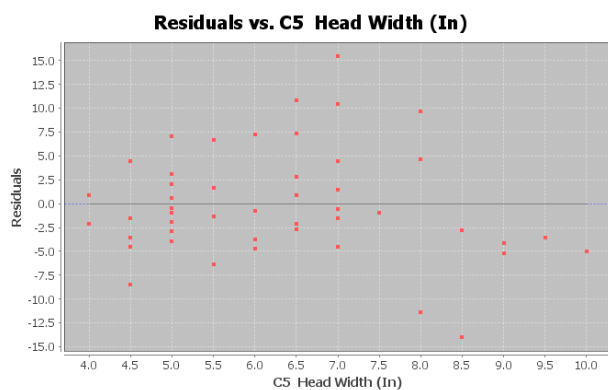
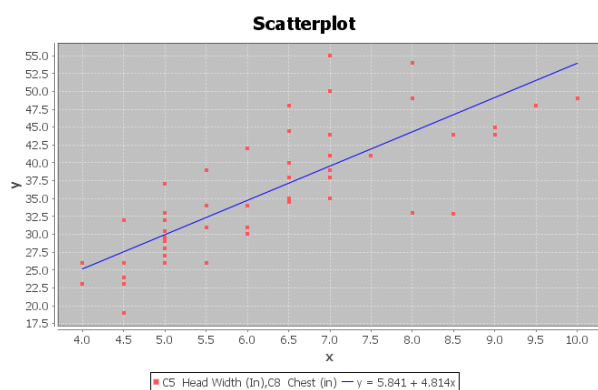
p-Value = $4.2208 \cdot 10^{-12}$

Regression:

Regression equation $Y = b_0 + b_1x$

$b_0 = 5.8408$

$b_1 = 4.8143$



29. Use a 5% significance level and the Statcato printout below to test the claim that there is NO relationship between the neck circumference (X) of a bear and its weight (Y). This printout came from the random bear data at www.matt-teachout.org.

Correlation and Regression: Significance level = 0.05

x = C6 Neck Circum (in)

y = C9 Weight (Lbs)

Sample size n = 54

Degrees of freedom = 52

	Test Statistic	Critical Value
r	0.9341	± 0.2681
t	18.8612	± 2.0067

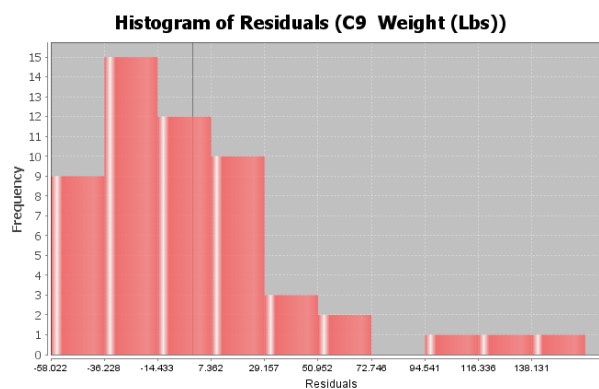
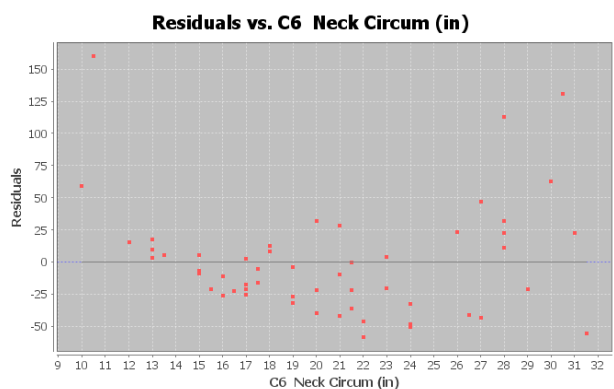
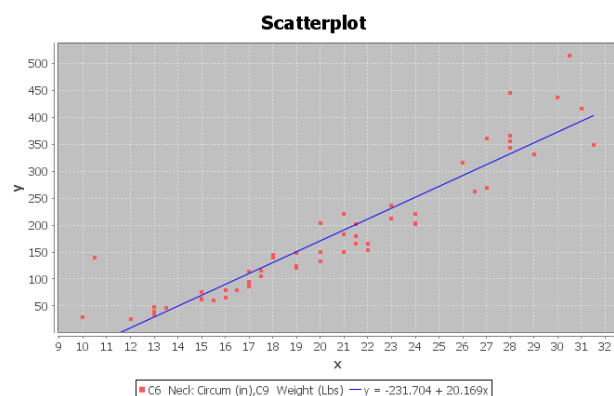
p-Value = 0

Regression:

Regression equation $Y = b_0 + b_1x$

$b_0 = -231.7044$

$b_1 = 20.1694$



(#30-32) Directions: Go to www.lock5stat.com and click on the “StatKey” button. Under “Randomization Hypothesis Tests”, click the one that says, “Test for Slope, Correlation”. Click “Generate 1000 Samples” a few times. Remember there are two ways of getting the P-value. If the top of the graph says “Randomization Dot plot of Correlation”, then the null hypothesis is $\rho = 0$. Remember rho looks like a “ ρ ” but it is not a “P”. If the top of the graph says

“Randomization Dot plot of Slope”, then the null hypothesis is $\beta_1 = 0$. Remember when StatKey simulates correlation we will be comparing the original r-value to all the simulated r-values in the simulation. When StatKey simulates the slope, we will be comparing the original sample slope to the simulated slopes. You will get about the same P-value from either of these. Assume the assumptions are met. Use the simulation in StatKey to answer the following questions.

- a) Write the null and alternative hypothesis for the correlation test. Address the quantitative relationship and label which is the claim. Is this a right-tailed, left-tailed, or two-tailed test?
- b) Write a sentence to explain the strength and direction based on the “original sample” correlation coefficient (r).
- c) Does the original sample correlation coefficient fall in a tail of the correlation simulation?
- d) Write a sentence to explain the original sample slope (b_1).
- e) Does the original sample slope fall in the tail of the slope simulation?
- f) Is the sample slope significantly different from zero? Explain your answer.
- g) Does the sample data significantly disagree with the null hypothesis? Explain your answer.
- h) Put the original sample slope into the slope simulation to calculate the P-value. What is your estimated P-value? (Answers will vary.)
- i) Write a sentence explaining the P-value.
- j) Compare the P-value to the significance level. Could the sample data or more extreme occur by sampling variability if the null hypothesis was true or is it unlikely? Explain your answer.
- k) Should we reject the null hypothesis or fail to reject the null hypothesis? Explain your answer.
- l) Write a conclusion for the test addressing evidence and the claim.
- m) Use the original sample slope, the estimated standard error in the simulation, and the following formula to calculate the T-test statistic. (Answers will vary.) Write a sentence to explain the T-test statistic.

$$T\text{-test statistic} = \frac{(\text{Slope} - 0)}{\text{Standard Error}}$$

30. Open the “Car Data” in Excel from www.matt-teachout.org. Copy and paste the miles per gallon (mpg) and horsepower into two columns in new excel spreadsheet. The mpg should be on the left and the horsepower should be on the right. The mpg will be the explanatory variable (X) and the horsepower will be the response variable (Y). Now go to www.lock5stat.com and click on StatKey. Under “Randomization Hypothesis Tests” click on “Test for Slope, Correlation”. Under “Edit Data” paste the two columns into StatKey. Now click “Generate 1000 Samples” a few times. Use the randomized simulation in StatKey and a 1% significance level to test the claim that there is a negative (inverse) relationship between mpg and horsepower.

31. Open the “Car Data” in Excel from www.matt-teachout.org. Copy and paste the horsepower and weight into two columns in new excel spreadsheet. The horsepower should be on the left and the weight should be on the right. The horsepower will be the explanatory variable (X) and the weight in tons will be the response variable (Y). Now go to www.lock5stat.com and click on StatKey. Under “Randomization Hypothesis Tests” click on “Test for Slope, Correlation”. Under “Edit Data” paste the two columns into StatKey. Now click “Generate 1000 Samples” a few times. Use the randomized simulation in StatKey and a 10% significance level to test the claim that there is a positive (direct) relationship between the horsepower and weight of a car.



32. Open the "Health Data" in Excel from www.matt-teachout.org. Copy and paste the age of women and the height of women into two columns in new excel spreadsheet. The age of women should be on the left and the height of women should be on the right. The age of women in years will be the explanatory variable (X) and the height of women in inches will be the response variable (Y). Now go to www.lock5stat.com and click on StatKey. Under "Randomization Hypothesis Tests" click on "Test for Slope, Correlation". Under "Edit Data" paste the two columns into StatKey. Now click "Generate 1000 Samples" a few times. Use the randomized simulation in StatKey and a 5% significance level to test the claim that there is NO relationship between the age and height of women.

Chapter 4 Review

1. Write down the definitions for the following key terms.

Correlation Coefficient (r) , R-squared , Standard Deviation of the Residual Errors , Slope , Residual , Y-Intercept , Explanatory Variable , Response Variable , Correlation , Regression , Scatterplot , Residual Plot , Regression Line , Histogram of the Residuals, Hypothesis Test , Sampling Variability (Random Chance) , P-value , Significance Level , Critical Value , Randomized Simulation , F-test statistic , Chi-Squared test statistic (χ^2) for the Goodness of Fit or Categorical Relationship Test, T-test statistic for correlation, Z-test statistic for two-population proportion test, T-test statistic for a two-population mean test

2. Write down the type of data and the null and alternative hypothesis for the following relationship hypothesis tests: Two-population proportion test, Goodness of Fit, Categorical Relationship Test, Two-population mean test, ANOVA, and Correlation.

3. Write down the assumptions for the following hypothesis tests: Two-population proportion test, Goodness of Fit, Categorical Association Test, Two-population mean test, ANOVA, and the Correlation Test.

4. Give the test statistic for each of the following hypothesis tests: Two-population proportion test, Goodness of Fit, Categorical Relationship Test, Two-population mean test, ANOVA, and Correlation.

5. Fill out the following table to interpret the given test statistics.

Test Statistic	Critical Value	Does the sample data significantly disagree with H_0 ?	Explain why.
$F = 2.174$	3.823		
$T = -2.556$	± 1.96		
$\chi^2 = 16.87$	9.977		
$F = 5.339$	2.742		
$T = 1.349$	± 2.576		
$\chi^2 = 1.883$	7.187		

6. Fill out the following table to interpret the given P-value.

P-value	P-value %	Significance Level	Does the sample data significantly disagree with H_0 ?	Could be random chance or Unlikely?	Reject H_0 or fail to reject?
0.238		5%			
0.0003		1%			
5.7×10^{-6}		10%			
0.441		5%			
0.138		1%			
0		10%			



7. Complete the table by writing the conclusions for the following.

P-value	Sig Level	Claim	Conclusion
0.238	5%	H_0	
0.0003	1%	H_A	
5.7×10^{-6}	10%	H_0	
0.441	5%	H_A	
0.138	1%	H_0	
0	10%	H_A	

8. If we want to see if two quantitative variables are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

9. If we want to see if two categorical variables with multiple options are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

10. If we want to see if categorical and quantitative variables are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

11. If we want to see if a categorical variables and a specific proportion are related, what hypothesis test should we use? What would the test statistic be? What assumptions should we check?

12. Suppose we want to see if the amount of money in peoples' checking accounts is related to city they live in. What hypothesis test should we use? Explain why.

13. Suppose we want to see if the percentage of people in a city that own an Android phone is related to the city they live. What hypothesis test should we use? Explain why.

14. Suppose we want to see if the amount of rainfall in areas across Europe is related to the number of fires in those areas. What hypothesis test should we use? Explain why.

15. Suppose we want to see if a person's type of health insurance is related to their education level. What hypothesis test should we use? Explain why.

16. An orthopedic surgeon that specializes in knee injuries is wondering if the proportion of knee injuries is the same for the various sports. (This would indicate that the percent of knee injuries is not related to the sport being played.) He looks through randomly selected knee injuries and finds the following data. What percentage of the knee injuries came from playing soccer? What percentage of the knee injuries came from playing tennis? What can these percentages tell us about the relationship? Since this data is looking at one proportion in six groups, what type of hypothesis test is this? Use the following Statcato printout and a 1% significance level to test the claim. Be sure to check expected values and the assumption necessary for the test. Give the chi-squared test statistic and the P-value, whether you reject the null hypothesis and a conclusion that the surgeon will understand. Write a sentence to explain the test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

Football	Baseball	Basketball	Soccer	Hockey	Tennis
23	8	14	31	19	5



Chi-Square Goodness-of-Fit Test:

Input: C4 observed counts

Expected frequency = 16.6667

Category	Observed Frequency	Expected Frequency	Contribution to χ^2
0	23.0	16.6667	2.4067
1	8.0	16.6667	4.5067
2	14.0	16.6667	0.4267
3	31.0	16.6667	12.3267
4	19.0	16.6667	0.3267
5	5.0	16.6667	8.1667

N	Number of Categories	DOF	Significance	Critical Value	Test statistics	p-Value
100.0	6	5	0.01	15.0863	28.16	$3.3869 \cdot 10^{-5}$

17. A forest ranger is looking into incidents of rabies among the animals. He thinks that the type of animal is related to whether or not they have rabies. What percentage of all the raccoons have rabies? What percentage of the squirrels have rabies? What percentage of the chipmunks have rabies? What can these probabilities show us about the relationship? This data was collected from one random sample of animals. From each animal, the type of animal was noted as well as their rabies status. Is this a Homogeneity test or an Independence test? Explain why. Use the following Statcato printout and a 5% significance level to test the claim that the type of animal is related to whether or not they have rabies. Be sure to check expected values and the assumption necessary for the test. Give the Chi-squared test statistic and the P-value, whether you reject the null hypothesis or fail to reject, and a conclusion that the ranger will understand. Write a sentence to explain the test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

	Squirrels	Chipmunks	Raccoons
Has Rabies	17	8	7
Does not have Rabies	21	22	20

Chi-Square Test: Contingency Table:

	Squirrels	Chipmunks	Raccoons	Total
Rabies	17.0 (12.8) [1.38]	8.0 (10.11) [0.44]	7.0 (9.09) [0.48]	32.0
No Rabies	21.0 (25.2) [0.70]	22.0 (19.89) [0.22]	20.0 (17.91) [0.25]	63.0
Total	38.0	30.0	27.0	95.0

(expected frequency), [test statistic contribution]

Significance Level	DOF	χ^2	Critical value	p-Value
0.05	2	3.4670	5.9915	0.1767



18. Go to www.lock5stat.com and click on the “StatKey” tab. Then click on “ χ^2 test for association”. On the top left part of the page click on “one true love by gender”. If it is not there, you can also click on “Edit Data” and type in the following contingency table. We want to determine if a person's belief about everyone having one true love is independent of (not related to) gender. What percent of the males believe that everyone has one true love? What percent of the females believe that everyone has one true love? What can these percentages tell us about the relationship? Use simulation and a 10% significance level to test the claim that gender and a person's belief about one true love are independent (not related). Be sure to check expected values and the assumption necessary for the test. Give the chi-squared test statistic and the simulated P-value, whether you reject the null hypothesis and a conclusion that the music students will understand. Write a sentence to explain the test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

[blank], Male, Female

Agree, 372, 363

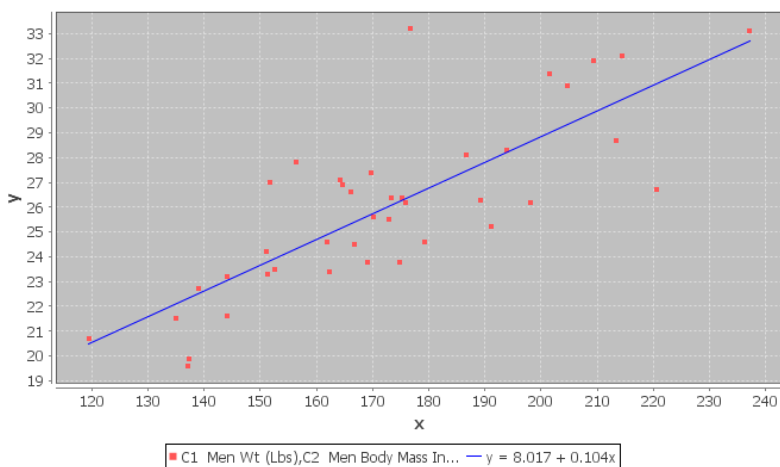
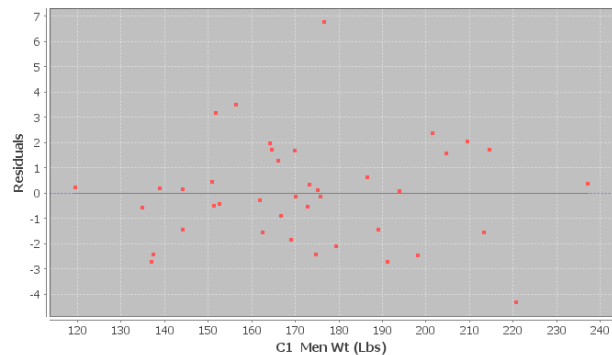
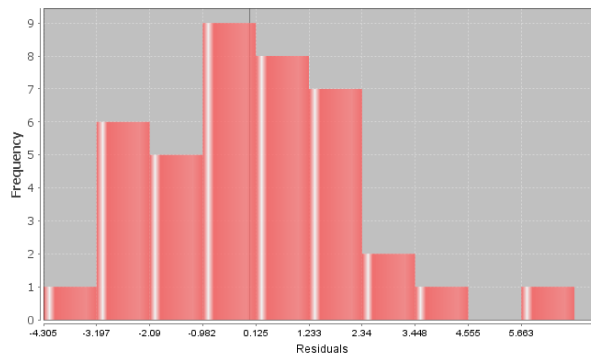
Disagree, 807, 1005

Don't Know, 34, 44

19. Go to www.matt-teachout.org and open the Math 140 Survey Fall 2019. Copy and paste the social media data and the money spent on meals data next to each other in a new Excel spreadsheet. The social media data should be on the left. Now copy the two columns together. Go to www.lock5stat.com and click on the “StatKey” button. Then click on “ANOVA for difference in means”. Under “Edit Data”, paste the two columns into StatKey and click “OK”. Use simulation and a 5% significance level to test the claim that a math 140 student's favorite social media is not related to how much they spend on meals. Be sure to check the assumptions, give the F-test statistic, null and alternative hypothesis, the simulated P-value, whether or not you reject the null hypothesis and a conclusion. Write a sentence to explain the F-test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.

20. We want to explore the relationship between the weight and body mass index for men. Let the explanatory variable (x) be the weight of the men and the response variable (y) be the body mass index (BMI) of the men. Write a sentence to explain the correlation coefficient r. Write a sentence to explain r-squared. Write two sentences to explain the two meanings of the standard deviation of the residual errors. Write a sentence to explain the meaning of the slope of the regression line. Write a sentence to explain the meaning of the y-intercept of the regression line. Use the regression line formula to predict the BMI of a man that weighs 185 pounds. How much error is there in that prediction? Use the following Statcato printout to perform a correlation hypothesis test with a 1% significance level to test the claim that there is a linear relationship between the weight of a man and his body mass index (BMI). Make sure to give make a scatterplot, residual plot, histogram of the residuals, the null and alternative hypothesis, the t-test statistic, the P-value, whether or not you reject the null hypothesis and a conclusion. Write a sentence to explain the t-test statistic. Write a sentence to explain the P-value. Was the sample data significant? Explain why. Could the sample data have happened by random chance or is it unlikely? Explain why.



Scatterplot**Residuals vs. C1 Men Wt (Lbs)****Histogram of Residuals (C2 Men Body Mass In...)**

Correlation and Regression: Significance level = 0.01

Series: C1 Men Wt (Lbs), C2 Men Body Mass In...

x = C1 Men Wt (Lbs)

y = C2 Men Body Mass In...

Sample size $n = 40$

Degrees of freedom = 38

Correlation:

$H_0: \rho = 0$ (no linear correlation)

$H_1: \rho \neq 0$ (linear correlation)

	Test Statistic	Critical Value
r	0.7997	± 0.4026
t	8.2095	± 2.7116

p-Value = $6.0619 \cdot 10^{-10}$

Regression:

Regression equation $Y = b_0 + b_1x$

$b_0 = 8.0169$

$b_1 = 0.1042$

Coefficient of determination $r^2 = 0.6395$

Standard Deviation of the Residual Errors = 2.0869



Appendix A Odd Answers

Section 1A Odd Answers

1. Bear Ages: Quantitative, Months
 Bear Month Data Taken: Categorical, 8 options (May-November)
 Bear Gender: Categorical, 2 options
 Head Length: Quantitative, Inches
 Head Width: Quantitative, Inches
 Neck Circumference: Quantitative, Inches
 Length: Quantitative, Inches
 Chest: Quantitative, Inches
 Weight: Quantitative, Pounds

3.
 - a) Milligrams of Aspirin: Quantitative
 - b) Types of Cars: Categorical
 - c) Smoke Marijuana or not: Categorical
 - d) Number of Bicycles: Quantitative
 - e) Types of Birds: Categorical
 - f) Grams of Gold: Quantitative
 - g) Types of Cardio Classes: Categorical
 - h) Number of Cardio Classes: Quantitative
 - i) City: Categorical
 - j) Money in Bank Accounts: Quantitative
 - k) Zip Codes: Categorical
 - l) Driver's License Numbers: Categorical
 - m) Number of Taxis: Quantitative

Section 1B Odd Answers

1. Population of Interest: All students at the college.
 Method: Voluntary Response
 Will not represent the population very well. There is sampling Bias, since the individuals were not chosen randomly.

3. Population of Interest: All students at the high school.
 Method: Convenience
 Will not represent the population very well. There is sampling Bias, since the individuals were not chosen randomly.

5. Population of Interest: All people in Rachael's home town.
 Method: Systematic
 Might represent the population. There is sampling bias, since the individuals were not chosen randomly, but it may be representative since the whole population was on the list. This data is not as biased as convenience or voluntary response, but not as good as a random sample.

7. Population of Interest: All employees at the company.
 Method: Census
 Census is better than a random sample. Will represent the population very well as long as there is no other types of bias present. No sampling bias.

9. Population of Interest: All people in Toronto.
 Method: Simple Random Sample
 Will represent the population well as long as there is no other types of bias present. No sampling bias.



11. Population of Interest: All people that use smart phones.
 Method: Voluntary Response
 Will not represent the population very well. Sampling bias, since the individuals were not chosen randomly.
13. Population of Interest: All teenagers and adults.
 Method: Stratified since they are comparing groups.
 Will represent the population well as long as there is no other types of bias present. Individuals were chosen randomly, so no sampling bias.
15. Population of Interest: All adults in North Carolina
 Method: Systematic and Convenience
 Will not represent the population very well. There is sampling bias since the individuals were not selected randomly. This data is particularly bad since most of the population has no opportunity to enter the store.
-

Section 1C Odd Answers

- 1.
- a) Population: All people or objects to be studied. For example, all students at College of the Canyons.
 - b) Census: Collecting data from everyone in your population. For example, collecting data from all of the students at college of the canyons.
 - c) Sample: Collecting data from a subgroup of the population. For example, collecting data from fifty students at College of the Canyons.
 - d) Bias: When data does not reflect the population. For example, friends and family will not represent the population of all people in Los Angeles, CA.
 - e) Question Bias: Phrasing a question in order to force people to answer the way you want. For example, we want to collect data on smoking cigarettes, but give the person a lecture on how unhealthy cigarettes are before asking them.
 - f) Response Bias: When someone is likely to lie about the answer to a question. For example, asking people how much they weigh in pounds. They may not give you a truthful answer.
 - g) Sampling Bias: Not using randomization when collecting sample data. For example, collecting data from only your friends and family. This is not a random sample.
 - h) Deliberate Bias: Falsifying or changing your data or leaving out groups from your population of interest. For example, a person might remove all of the data from people that disagreed with their opinion.
 - i) Non-response Bias: When people are likely to not answer when asked to provide data. Randomly calling phone numbers to get data, but the person refuses to answer the phone.
3. Population of interest: All people in the U.S.
- Question Bias: The question was phrased to make people feel bad about answering no.
 Response Bias: Vaccinations are a controversial issue and many people may feel scared to admit that they don't agree with vaccinations.
 Non-response Bias: There will be many people that randomly selected, but refuse to answer the question.
5. Population of interest: All Americans.
- Response Bias: Cocaine users would not feel comfortable answering the question honestly.
 Non-response: Many people may be randomly selected, but will chose not to answer the question.
7. Population of interest: All adults in Palmdale, CA.
- Sampling Bias: The individuals were not selected randomly.
 Deliberate Bias: Julie skipped streets that looked poor. These people are not being represented in the data.
 Response Bias: People often lie about their income.
 Non-response: Many people may not be home or refuse to answer the door.



9. Population of interest: All pills made by the company.

Deliberate Bias: They deleted data that poorly reflected the pharmaceutical company.

Section 1D Odd Answers

1. Explanatory Variable: Having a cell phone or not.

Response Variable: Ruler catch length in inches or “drop”.

2. We needed a control group the measures the classes ability to catch a ruler in general. We can then compare the cell phone data to the control group.

3. The two groups were people with the cell phone (treatment group) and those without a cell phone (control group.) They were perfectly alike in all confounding variables since they were the same exact people measured twice.

4. Answers may vary. Confounding Variables: Age, hand-eye coordination, distractions besides the phone, hand size, ability to text one handed, position of the ruler when dropped, ...

Since the same people were measured twice, the two groups had the exact same ages, hand-eye coordination and ability to text one-handed. The amount of distraction was relatively the same with or without the phone. The instructor gave a demonstration so that everyone would hold and drop the ruler the same way.

5. Neither. The explanatory variable was having a cell phone or not. The person knew whether they had a cell phone or not. Not knowing when the ruler would be dropped does not constitute blind since it is the response variable.

6. Answers will vary from class to class. The average catch distance was lower for the no cell phone group, but it is difficult to determine if it is significant at this point. We will learn that later. The number of drops was significantly greater in the cell phone group. Since confounding variables were controlled, we have proven that texting does cause you to drop the ruler more often. Whether this experiment applies to texting while driving is debatable. Some people have said that dropping the ruler may be equivalent to not hitting the breaks in time. Again, this is debatable. The experiment does prove that texting slows reflexes and you do need reflexes when you drive.

7. Observational Study: Collecting data without trying to control confounding variables. Data collected by an observational study can show relationships but cannot prove cause and effect.

8. Experiment: A scientific method for controlling confounding variables and proving cause and effect.

9. Explanatory Variable: The independent or treatment variable. In an experiment, this is the variable that causes the effect.

10. Response Variable: The dependent variable. In an experiment this the variable that measures the effect.

11. Confounding Variables (or lurking variables): Other variables that might influence the response variable other than the explanatory variable being studied.

12. Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

13. Placebo: A fake medicine or fake treatment used to control the placebo effect.

14. Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

15. Single Blind: When only the person receiving the treatment does not know if it is real or a placebo.

16. Double Blind: When both the person receiving the treatment and the person giving the treatment does not know if it is real or a placebo.



17. This is an experiment, since they need to prove that the medicine has the desired effect. They must control confounding variables like the amount of motion, genetics, age, diet, pregnancy, etc. If they control all of the confounding variables and the medicine (treatment) group has significantly less motion sickness, they will have succeeded in proving cause and effect.

19. This is an observational study since they just collected data without thought to controlling confounding variables. This can show the number of cases of tuberculosis is related to or associated with the low income, crowded cities, but it will not be able to prove cause and effect.

21. This is an observational study since they just collected data without thought to controlling confounding variables. This can show that obesity is related to or associated with having diabetes, but they will not be able to prove cause and effect. There are many variables involved in determining why someone has diabetes.

Section 1E Odd Answers

1.

- a) 0.75
- c) 0.00664
- e) 0.397
- g) 0.00189
- i) 0.0316
- k) 0.961
- m) 0.00007
- o) 0.662
- q) 1

2.

- a) 5.7%
- c) 0.33%
- e) 6.13%
- g) 0.045%
- i) 4.6%
- k) 0.27%
- m) 0.58%
- o) 100%
- q) 2.04%

3.

$$15\% = 0.15$$

$$\text{Estimated Amount} = 0.15 \times 78300 = 11,745$$

We estimate that approximately 11,745 people in Chino Hills are without health insurance.

5.

$$9.3\% = 0.093$$

$$\text{Estimated Amount} = 0.093 \times 18400 = 1711.2 \approx 1711$$

We estimate that approximately 1,711 students at COC have diabetes.



7.

$$1.47\% = 0.0147$$

$$\text{Estimated Amount} = 0.0147 \times 136400 = 2005.08 \approx 2005$$

We estimate that approximately 2,005 people in Van Nuys have autism.

9.

$$14.8\% = 0.148$$

$$\text{Estimated Amount} = 0.148 \times 305700 = 45243.6 \approx 45244$$

We estimate that approximately 45,244 people in Stockton live below the poverty line.

11.

$$\text{Athletic Wear: } 139/213 \approx 0.653 = 65.3\%$$

$$\text{Traditional Jeans: } 74/213 \approx 0.347 = 34.7\%$$

$$\text{Percent of Increase} = (0.653 - 0.347)/0.347 = 0.882 = 88.2\% \text{ of increase.}$$

The percent of women that prefer athletic wear does seem to be significantly higher than the percent that prefer traditional jeans. It is also practically significant since there was 65 more women in the sample that preferred athletic wear.

13.

$$\text{Med/Surg: } 57/350 \approx 0.163 = 16.3\%$$

$$\text{Telemetry: } 49/350 \approx 0.14 = 14\%$$

$$\text{Percent of Increase} = (0.163 - 0.14)/0.14 \approx 0.164 = 16.4\% \text{ of increase.}$$

The percent of patients admitted to telemetry and med/surge seem very close. The percent of increase is very small. There were also only 8 more patients in Med/Surg than telemetry. These indicate there is no significant difference, practically or statistically.

15.

$$\text{Medicine: } 13/57 \approx 0.228 = 22.8\%$$

$$\text{Placebo: } 11/61 \approx 0.180 = 18.0\%$$

$$\text{Percent of Increase} = (0.228 - 0.18)/0.18 \approx 0.267 = 26.7\% \text{ of increase.}$$

The percent of patients that improved on the medicine is only slightly higher than the placebo group. Practically there is not much difference. Only two more patients on the medicine showed improvement. That is not practically significant.

17.

$$\text{Proportion of female} \approx 0.591$$

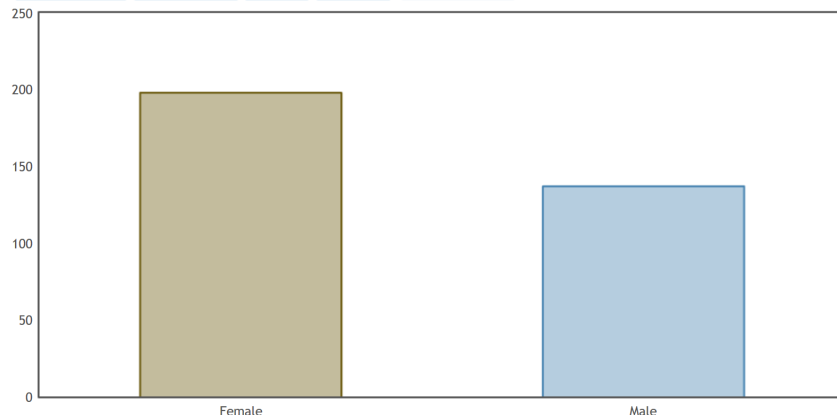
$$\text{Proportion of male} \approx 0.409$$

Percent of Increase = $(0.591 - 0.409)/0.409 \approx 0.445 = 44.5\%$ of increase. This indicates that the percentage of female COC statistics students is significantly higher than the percentage of male students. It is also practically significant since there were 61 more female students than male.



StatKey Descriptive Statistics for One Categorical Variable

Custom Dataset Show Data Table Edit Data Upload File Change Column(s)


Summary Statistics

	Count	Proportion
Female	198	0.591
Male	137	0.409
Total	335	1.000

19.

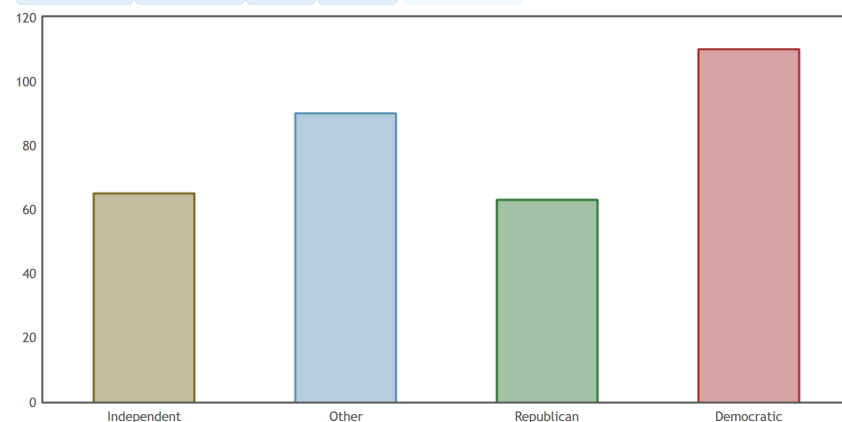
 Proportion of Democrat ≈ 0.335

 Proportion of Republican ≈ 0.192

Percent of Increase = $(0.335 - 0.192)/0.192 \approx 0.745 = 74.5\%$ of increase. This indicates that the percentage of COC statistics students that are democrat is significantly higher than republican. It is practically significant also since there are 47 more democratic statistics students than republican.

StatKey Descriptive Statistics for One Categorical Variable

Custom Dataset Show Data Table Edit Data Upload File Change Column(s)


Summary Statistics

	Count	Proportion
Independent	65	0.198
Other	90	0.274
Republican	63	0.192
Democratic	110	0.335
Total	328	1.000

21.

 Percentage of cars made in France $\approx 3\%$

Number of cars made in U.S. = 22

 Proportion of cars made in Sweden ≈ 0.05

 Proportion of cars made in Japan ≈ 0.18

 Proportion of cars made in Germany ≈ 0.13

 Percent of Increase (Japan & Germany) $\approx (0.18 - 0.13) / 0.13 \approx 0.385 = 38.5\%$


The percent of increase does seem to indicate that the percent of cars made in Japan is significantly higher than Germany. However does not seem to be practically significant since the number of cars made in Japan was only 2 more than Germany. Overall, I would say it is not significant.

23.

Percentage of cereals made by Quaker $\approx 17\%$

Number of cereals made by Ralston = 3

Proportion of cereals made by General ≈ 0.29

Proportion of cereals made by Kelloggs ≈ 0.33

Proportion of cereals made by Quaker ≈ 0.17

Percent of Increase (Kelloggs & Quaker) $\approx (0.33 - 0.17) / 0.17 \approx 0.941 = 94.1\%$

The percent of increase indicates that the percent of cereals made by Kelloggs is significantly higher than Quaker. However does not seem to be practically significant since it was a small sample size and the number of cereals made by Kelloggs was only 4 more than Quaker.

25.

Percentage of cereals on top shelf $\approx 33\%$

Number of cereals on bottom shelf = 8

Proportion of cereals on middle shelf ≈ 0.33

Percent of Increase (top and bottom shelves) $\approx (0.33 - 0.33) / 0.33 \approx 0 = 0\%$

There is no significant difference between the percentages of cereals put on the top and bottom shelves. They appear to be about the same.

27.

a) $0.122 = 12.2\%$

b) $0.113 = 11.3\%$

c) $0.796 = 79.6\%$

d) $1 - 0.796 = 0.204 = 20.4\%$

e) $0.110 = 11.0\%$

f) $1 - 0.110 = 0.89 = 89\%$



Binomial Distribution: $n=84$, $p=0.12$

Input: 11.0

Type: Probability density

X $P(X)$

11.0 0.122219

Binomial Distribution: $n=84$, $p=0.12$

Input: 8.0

Type: Probability density

X $P(X)$

8.0 0.113128

Binomial Distribution: $n=84$, $p=0.12$

Input: 12.0

Type: Cumulative probability

X $P(\leq X)$

12.0 0.796146

Binomial Distribution: $n=84$, $p=0.12$

Input: 6.0

Type: Cumulative probability

X $P(\leq X)$

6.0 0.109721

29.

a) $0 = 0\%$ b) $1 - 0 = 1 = 100\%$ **Binomial Distribution: $n=57$, $p=0.845$**

Input: 9.0

Type: Cumulative probability

X $P(\leq X)$

9.0 0

Section 1F Odd Answers

1.

a) The shape of the data is relatively bell shaped or unimodal and symmetric. In a histogram, the tallest bars are relatively in the middle and the left and right tail are about the same length.

b) The mean average is the average that we use when the data is normal. It also balances the distances. It is calculated by adding all of the data values and dividing the sum by the sample size n .



c) The standard deviation calculates the average distance that data values are from the mean. It is the measure of spread used when data is normal. To calculate the standard deviation for one quantitative data set, take every number in the data set and subtract the mean from it. Then square the differences. Add up all the squares and divide by $n - 1$ where n is the sample size. Now take the square root of the answer. Always have a computer calculate the standard deviation for you.

2.

- a) Mean Average
- b) Standard Deviation
- c) One Standard Deviation or less.
- d) Mean \pm Standard Deviation
- e) 68%
- f) Two or more standard deviations.
- g) 2.5%
- h) 2.5%

3.

- a) The data measures the neck circumference of bears. The units are inches.
- b) 54 total bears
- c) Yes. The data is nearly normal (almost bell shaped).
- d) 10 inches
- e) 31.5 inches
- f) Center = Mean Average = 20.556 inches
- g) Typical Spread = Standard Deviation = 5.641 inches
- h)

$$20.556 - 5.641 = 14.915 \text{ inches}$$

$$20.556 + 5.641 = 26.197 \text{ inches}$$

Typical bears have a neck circumference between 14.915 inches and 26.197 inches.

i)

$$\text{Unusual High Cutoff: } 20.556 + (2 \times 5.641) = 20.556 + 11.282 = 31.838 \text{ inches.}$$

Any neck circumference of 31.838 inches or more would be considered unusually high (high outlier).

j)

$$\text{Unusual Low Cutoff: } 20.556 - (2 \times 5.641) = 20.556 - 11.282 = 9.274 \text{ inches.}$$

Any neck circumference of 9.274 inches or less would be considered unusually low (low outlier).

k)

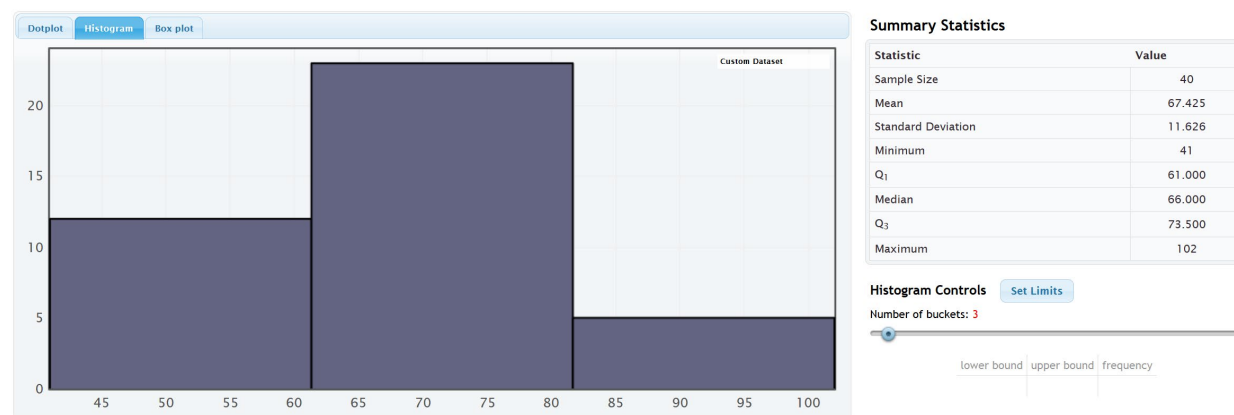
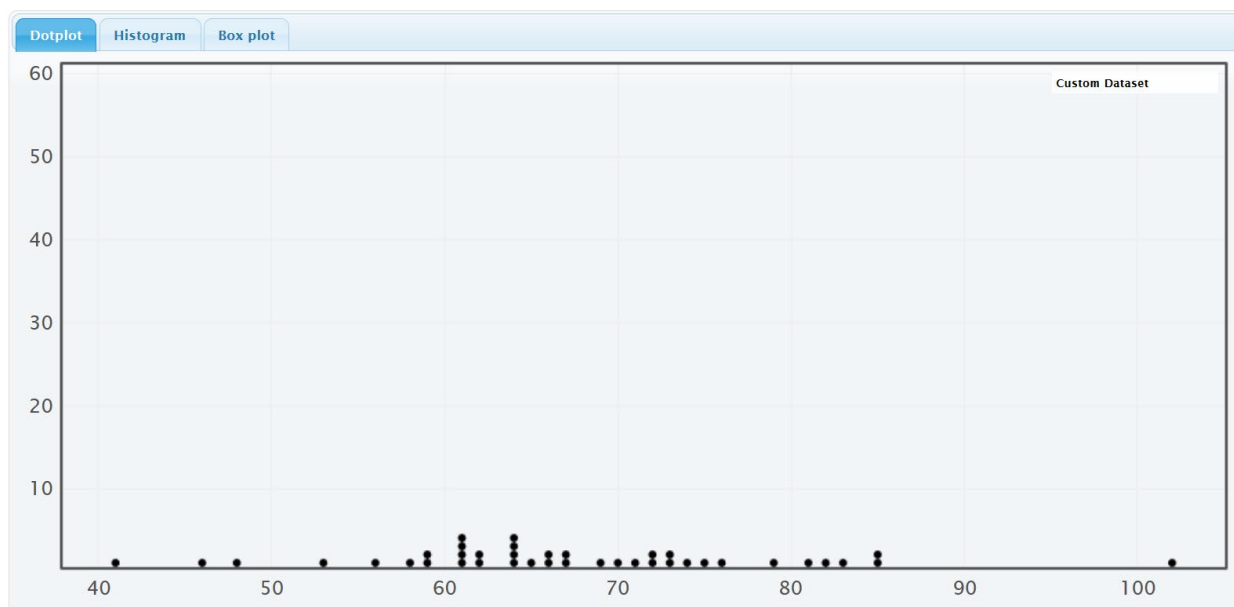
There are no unusually large bear neck sizes. The largest neck size is 31.5 inches which is not higher than the unusually high cutoff of 31.838 inches.

l)



There are no unusually small bear neck sizes. The smallest neck size is 10 inches which is not lower than the unusually low cutoff of 9.274 inches.

5.



a) The data is measuring the diastolic blood pressure of women. The units are millimeters of mercury (mm of Hg).

b) There were 40 total women in the data set.

c) The data is nearly normal (almost bell shaped).

d) Min = 41 mm of Hg

e) Max = 102 mm of Hg

f) Center = Mean Average = 67.425 mm of Hg

g) Typical Spread = Standard Deviation = 11.626 mm of Hg

h)

Mean – Standard Deviation = $67.425 - 11.626 = 55.799$ mm of Hg

Mean + Standard Deviation = $67.425 + 11.626 = 79.051$ mm of Hg



Typical women in this data have a diastolic blood pressure between 79.051 mm of Hg and 55.799 mm of Hg.

i)

Unusual High Cutoff: $67.425 + (2 \times 11.626) = 67.425 + 23.252 = 90.677$ mm of Hg.

Any woman in the data with a diastolic blood pressure of 90.677 mm of Hg or higher would be considered unusually high (high outlier).

j)

Unusual Low Cutoff: $67.425 - (2 \times 11.626) = 67.425 - 23.252 = 44.173$ mm of Hg.

Any woman in the data with a diastolic blood pressure of 44.173 mm of Hg or lower would be considered unusually low (low outlier).

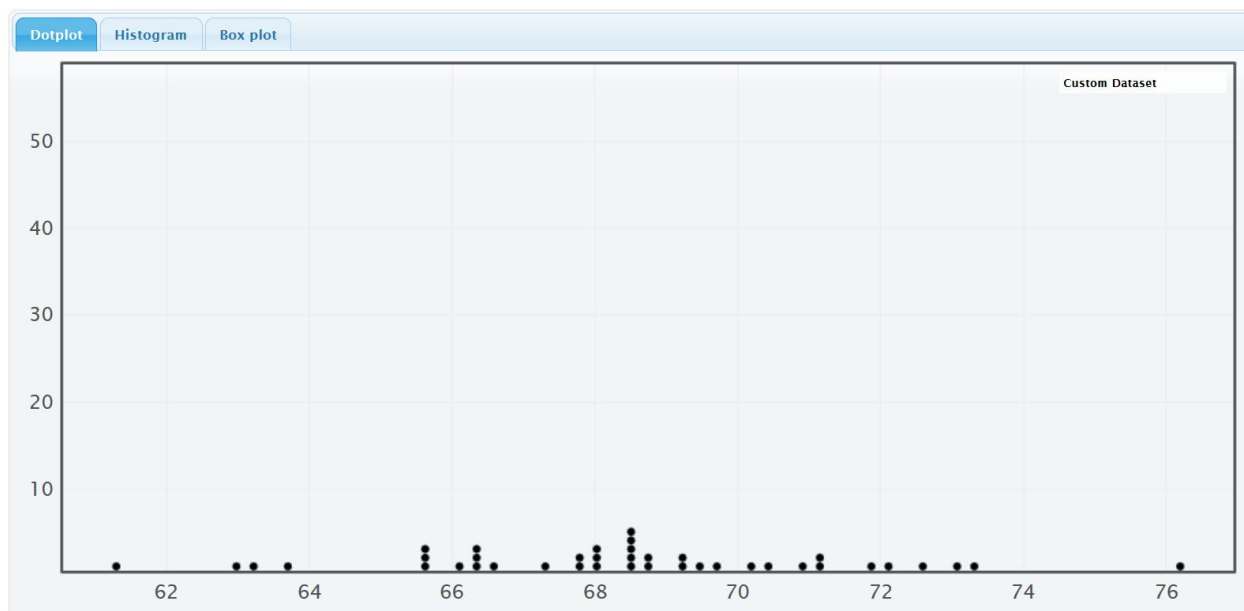
k)

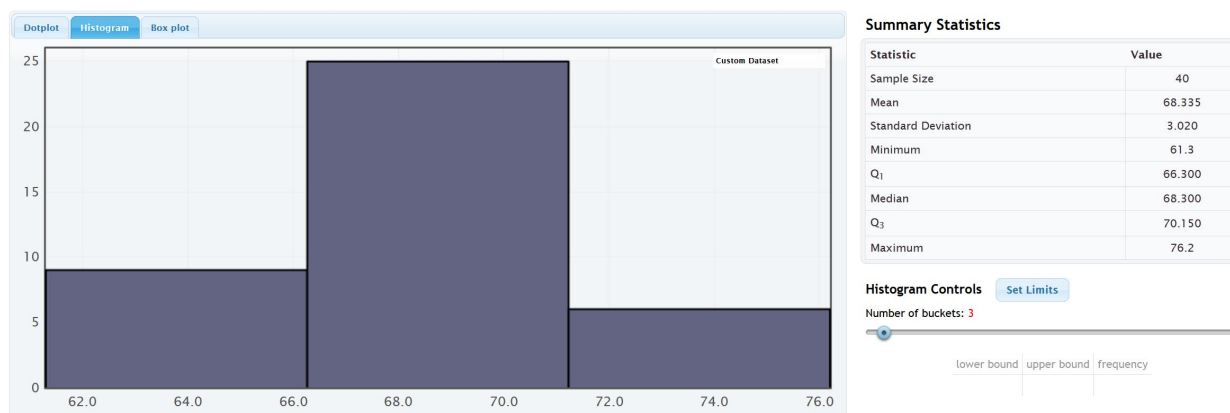
The dot plot shows only one dot above the unusual high cutoff of 90.677 mm of Hg. It is the maximum value of 102 mm of Hg. So the woman with a diastolic blood pressure of 102 mm of Hg is considered unusually high or a high outlier.

l)

The dot plot shows only one dot below the unusual low cutoff of 44.173 mm of Hg. It is the minimum value of 41 mm of Hg. So the woman with a diastolic blood pressure of 41 mm of Hg is considered unusually low or a low outlier.

7.





a) The data is measuring the heights of men. The units are inches.

b) There were 40 total men in the data set.

c) The data is nearly normal (almost bell shaped).

d) Min = 61.3 inches

e) Max = 76.2 inches

f) Center = Mean Average = 68.335 inches

g) Typical Spread = Standard Deviation = 3.020 inches

h)

Mean – Standard Deviation = $68.335 - 3.020 = 65.315$ inches

Mean + Standard Deviation = $68.335 + 3.020 = 71.355$ inches

Typical men in this data have a height between 65.315 inches and 71.355 inches.

i)

Unusual High Cutoff: $68.335 + (2 \times 3.020) = 68.335 + 6.040 = 74.375$ inches.

Any man in the data with a height of 74.375 inches or more would be considered unusually high (high outlier).

j)

Unusual High Cutoff: $68.335 - (2 \times 3.020) = 68.335 - 6.040 = 62.295$ inches.

Any man in the data with a height of 62.295 inches or less would be considered unusually low (low outlier).

k)

The dot plot shows only one dot above the unusual high cutoff of 74.375 inches. It is the maximum value of 76.2 inches. So the man with a height of 76.2 inches is considered unusually tall or a high outlier.

l)

The dot plot shows only one dot below the unusual low cutoff of 62.295 inches. It is the minimum value of 61.3 inches. The next largest dot was 62.9 inches which is not below the cutoff. So the man with a height of 61.3 inches is considered unusually short or a low outlier.



9.

A Z-score is the number of standard deviations that a value is from the mean.

10.

Typical values have a Z-score between -1 and $+1$ inclusively.

11.

For normal data, a Z-score of $+2$ or higher would indicate that the data value is unusually high or a high outlier.

For normal data, a Z-score of -2 or less would indicate that the data value is unusually low or a low outlier.

13.

a) $Z = (89 - 99.8) / 15.3 \approx -0.71$

b) Jan's IQ score was only 0.71 standard deviations below the mean.

c) Jan's IQ is not unusual, since the Z-score is not $+2$ or above or -2 or below. In fact, she has a very typical IQ, since the Z-score since it is between -1 and $+1$.

15.

a) $Z = (13.61 - 46.89) / 12.44 \approx -2.68$

b) The amount of money that Julie spent is 2.68 standard deviations below the mean.

c) The amount that Julie spent was unusually low (low outlier) since the Z-score was below -2 .

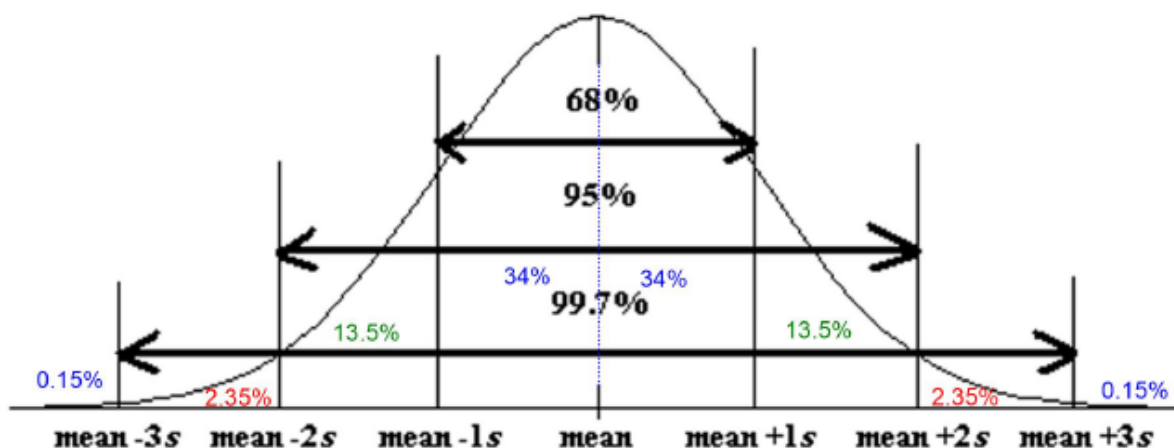
17.

a) $Z = (57 - 35.663) / 9.352 \approx +2.28$

b) This bears chest size is 2.28 standard deviations above the mean.

c) This bears chest size is unusually large (high outlier) since the Z-score is greater than $+2$.

19.



21.

a) $34\% + 34\% + 13.5\% = 81.5\%$

b) $34\% + 34\% + 13.5\% + 2.35\% + 0.15\% = 84\%$

c) One standard deviation from the mean is typical. So typical bear neck circumferences are between 14.915 inches and 26.197 inches.

d) The unusual high cutoff is two standard deviations above the mean which is 31.838 inches. So any bear with a neck circumference of 31.838 inches or more is considered unusually large.

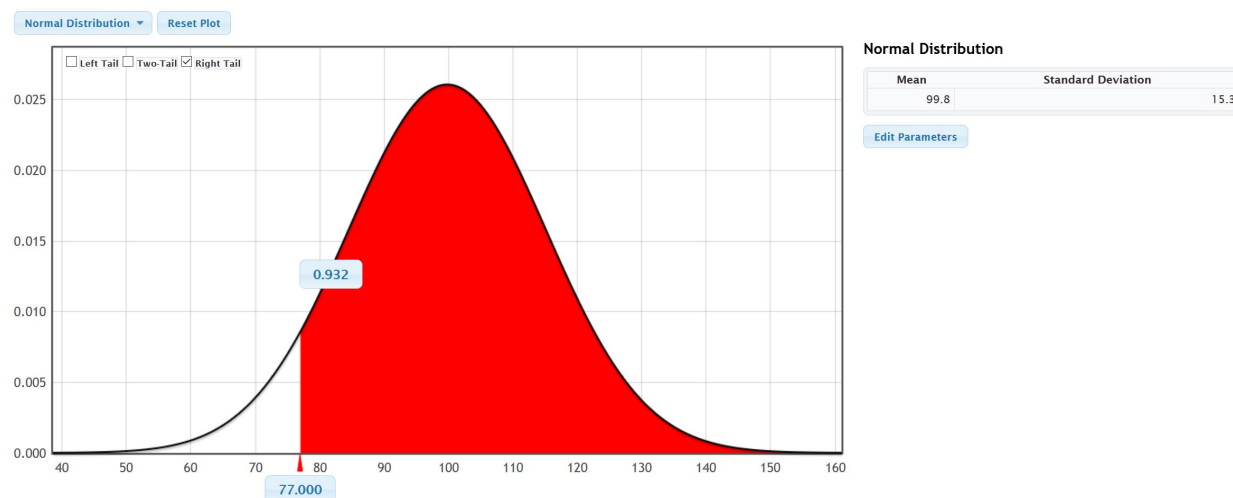
e) The unusual low cutoff is two standard deviations below the mean which is 9.274 inches. So any bear with a neck circumference of 9.274 inches or less is considered unusually small.

f) One standard deviation below the mean has 84% above. So 84% of the bears have a neck circumference greater than 14.915 inches.

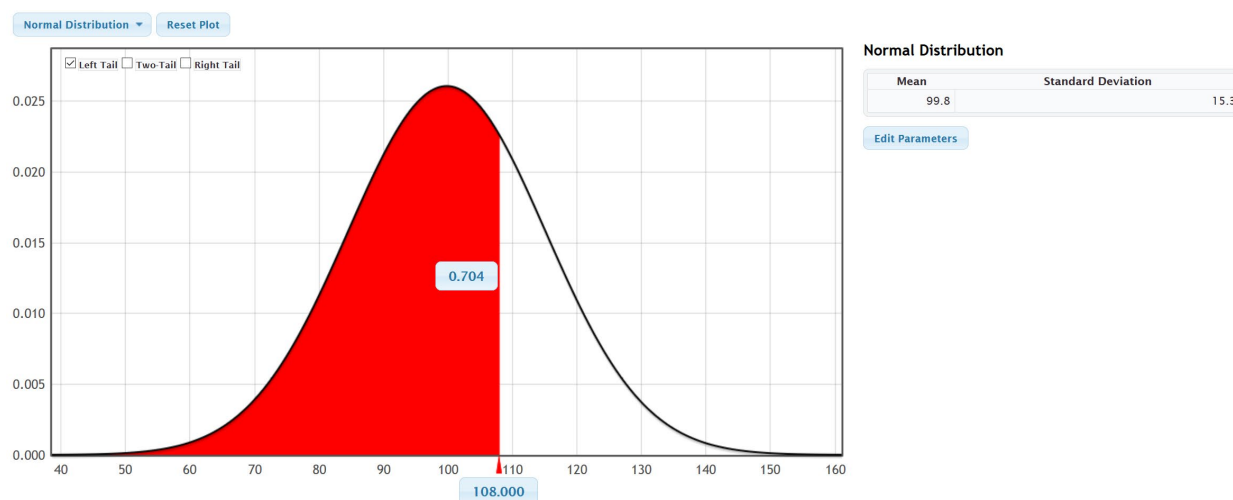
g) $13.5\% + 2.35\% + 0.15\% = 16\%$

23.

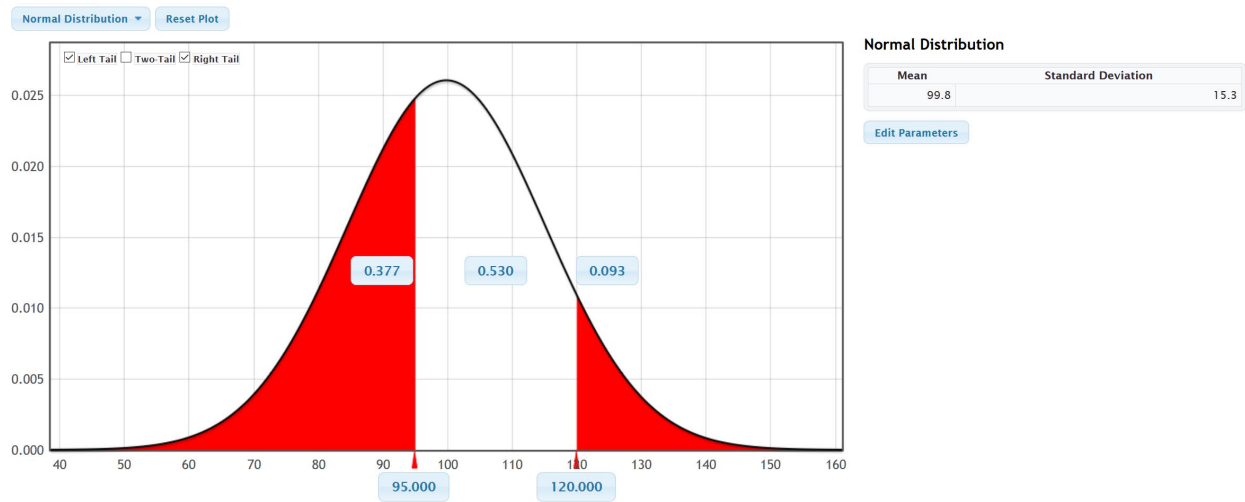
a) Based on this data, about 93.2% of people have an IQ above 77.



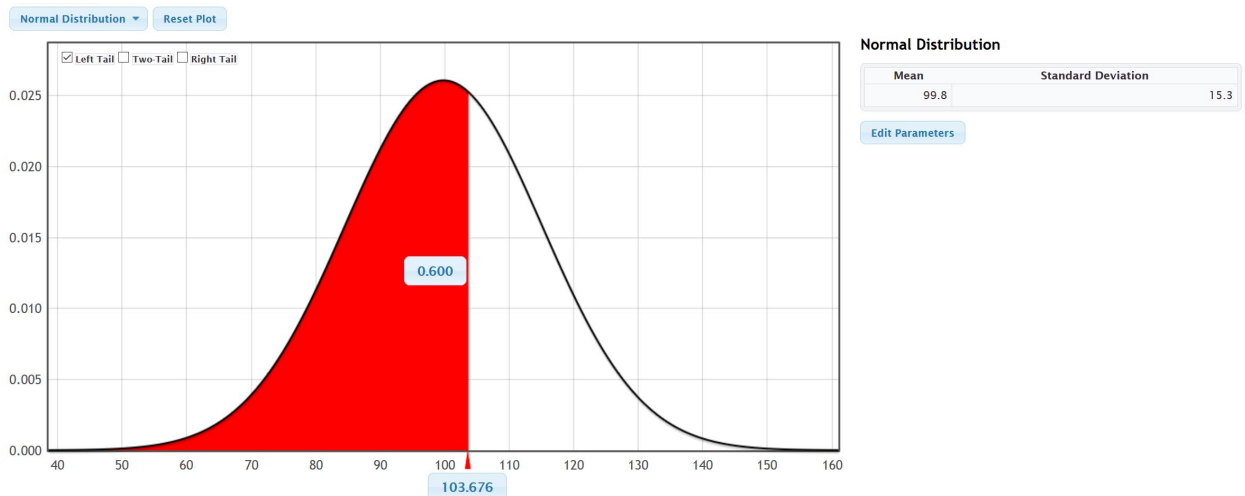
b) Based on this data, about 70.4% of people have an IQ below 108.



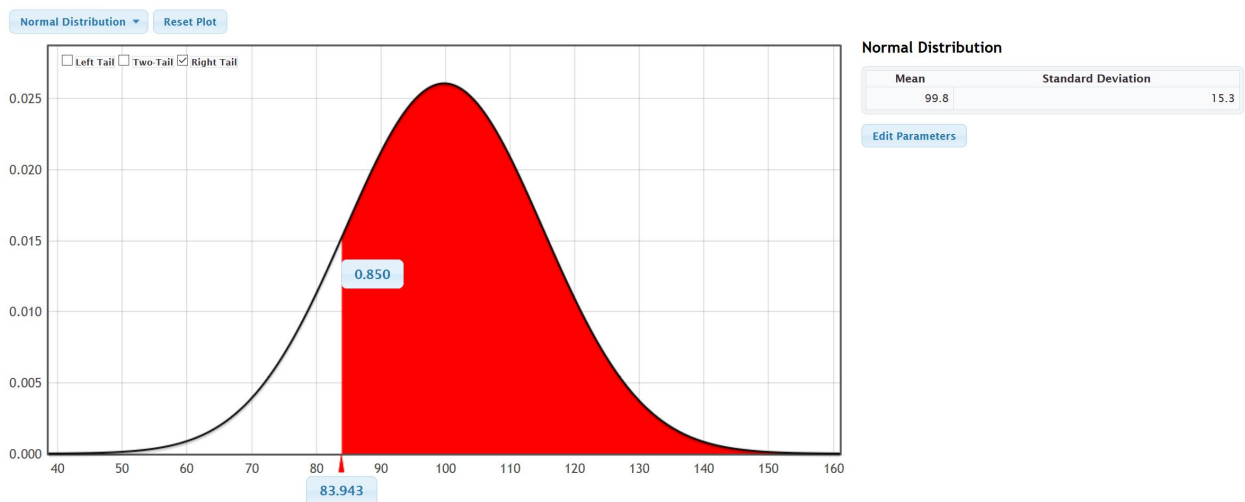
c) Based on this data, about 53.0% of people have an IQ between 95 and 120.



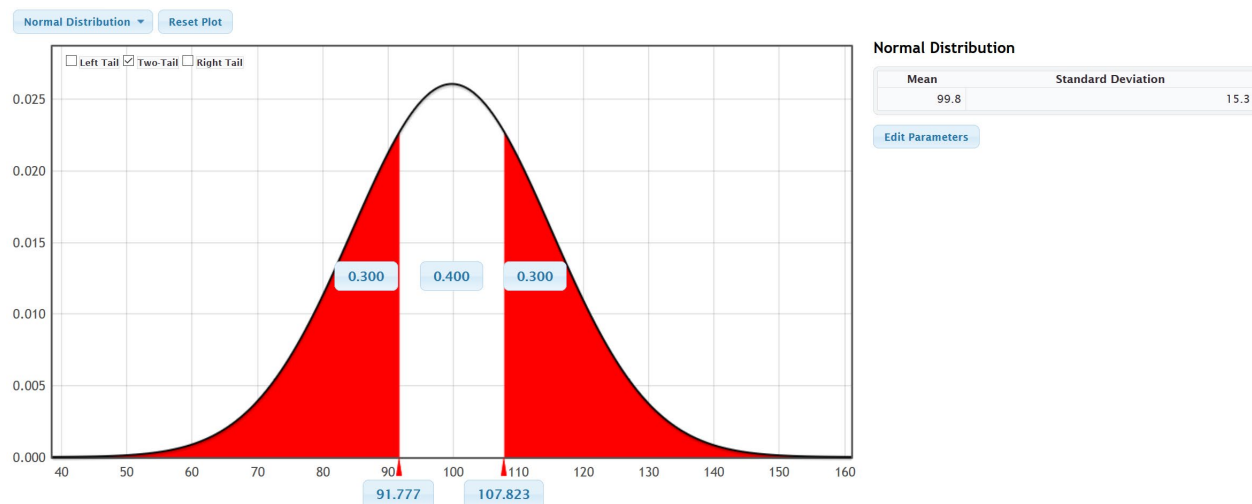
d) About 60% of people have an IQ below 103.676.



e) Based on this data, about 85% of people have an IQ above 83.943.

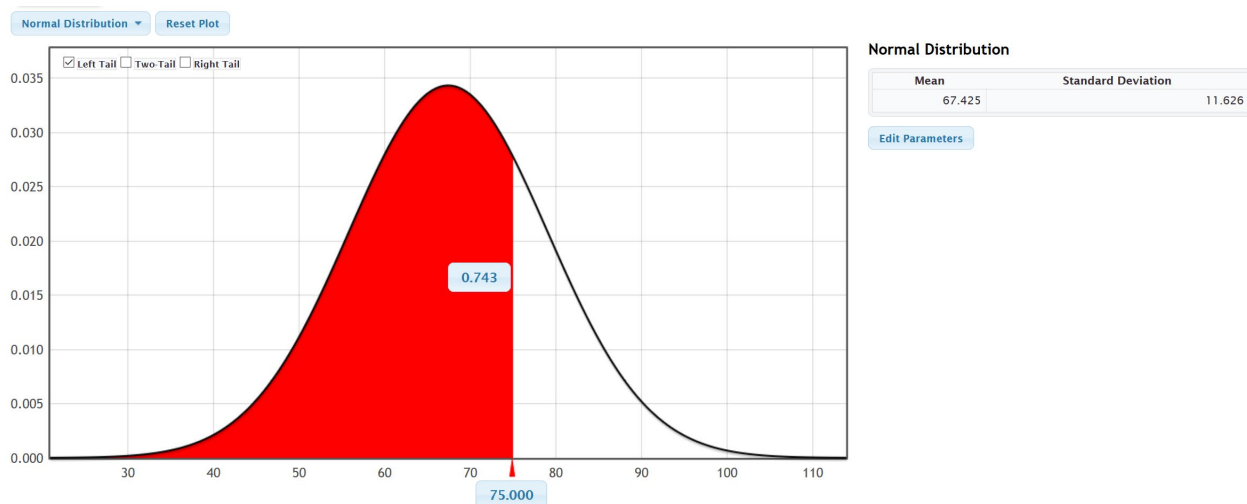


f) Based on this data, the middle 40% of people have an IQ between 91.777 and 107.823.

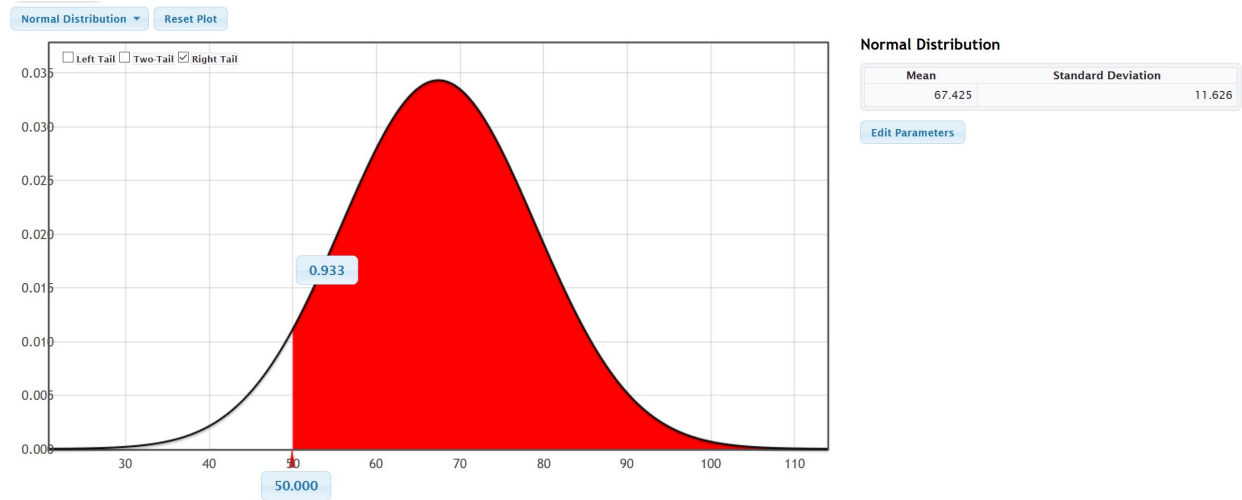


25.

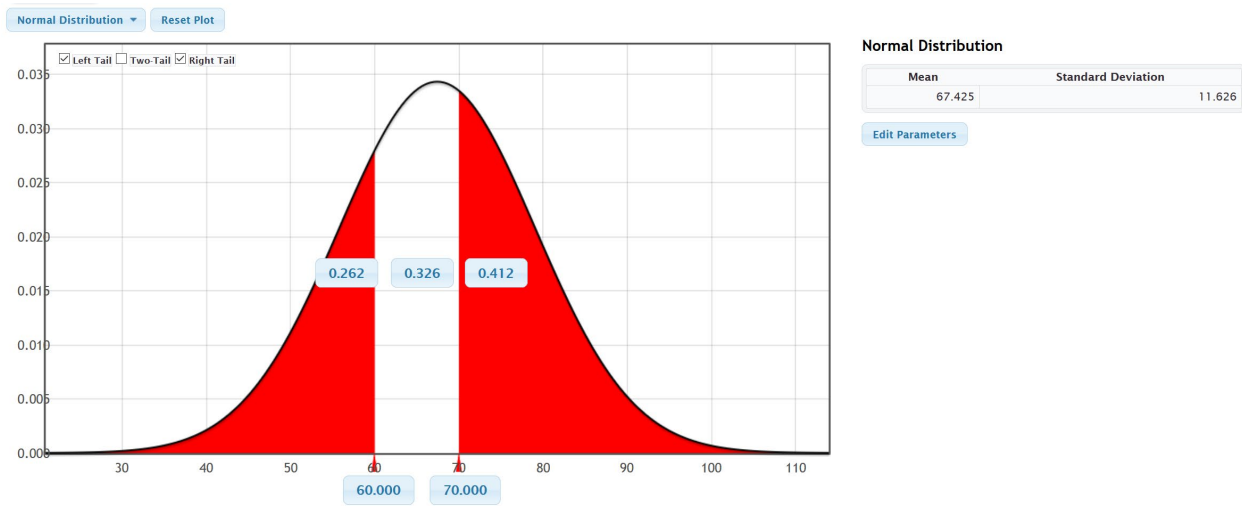
a) Based on this data, about 74.3% of women have a diastolic blood pressure below 75 mm of Hg.



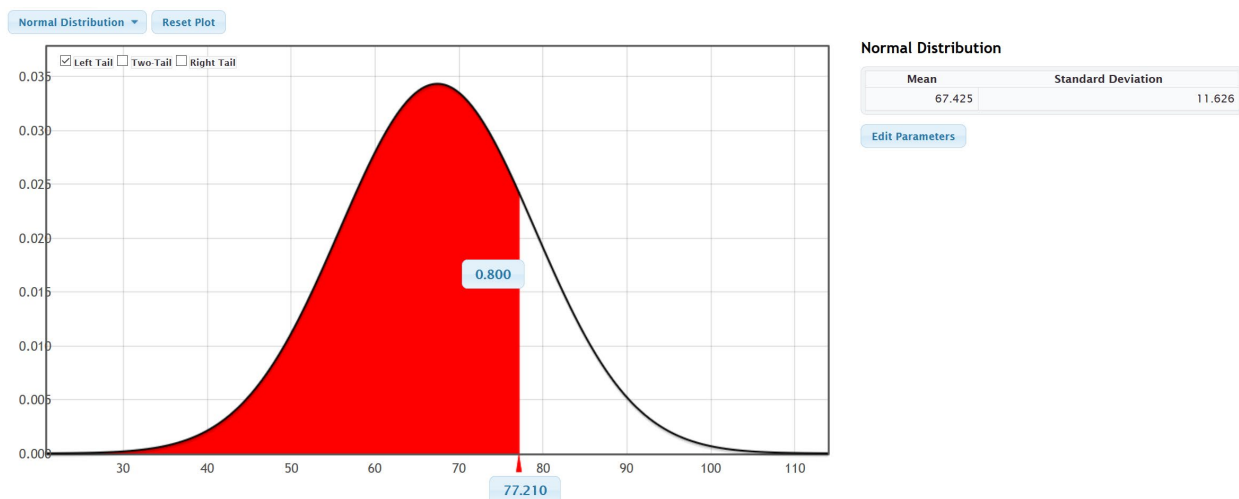
b) Based on this data, about 93.3% of women have a diastolic blood pressure above 50.



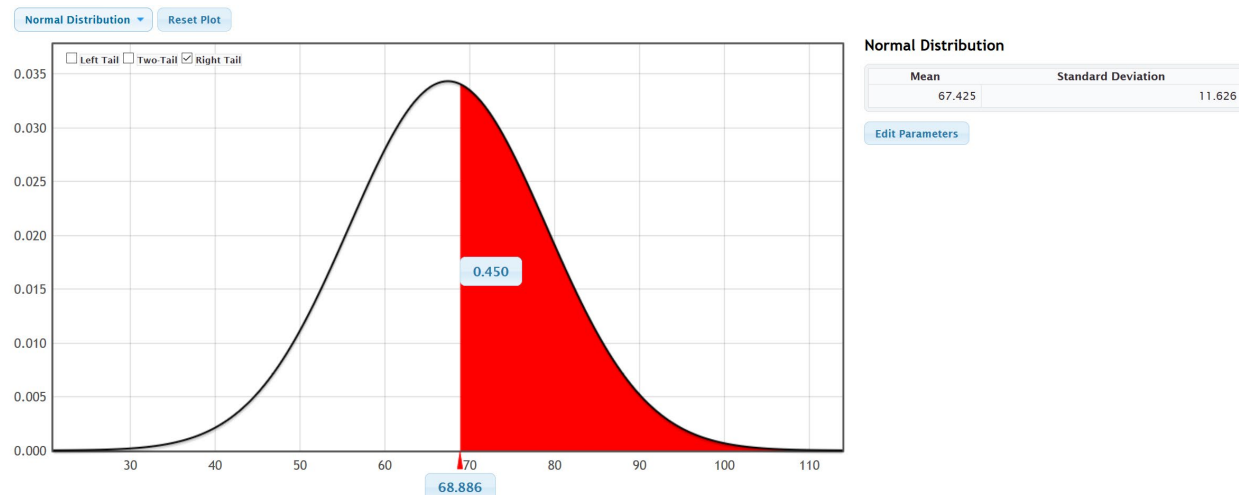
c) Based on this data, about 32.6% of women have a diastolic blood pressure between 60 and 70.



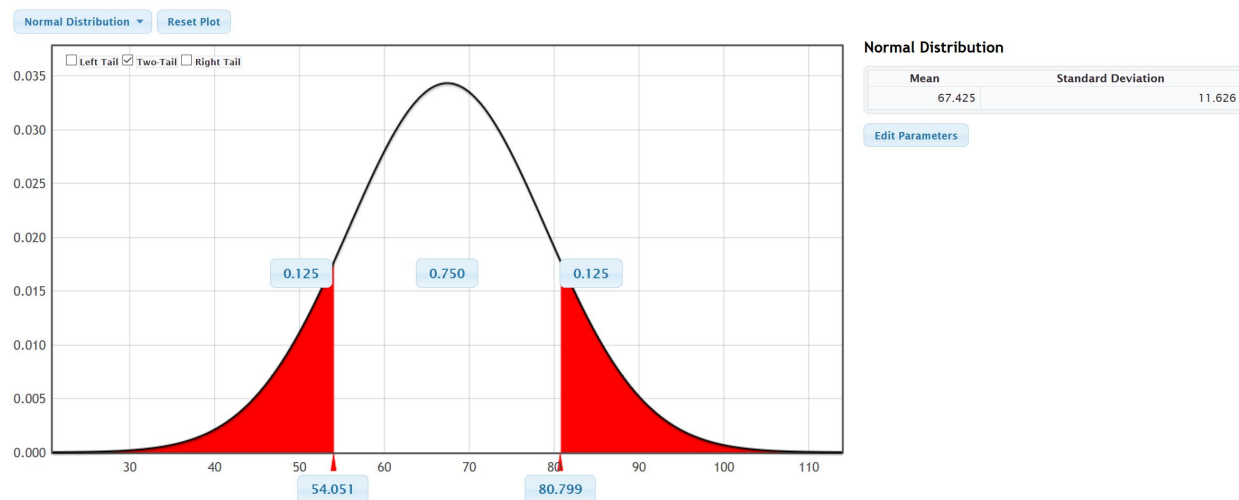
d) Based on this data, about 80% of women have a diastolic blood pressure less than 77.21 mm of Hg.



e) Based on this data, about 45% of women have a diastolic blood pressure above 68.886 mm of Hg.



f) Based on this data, the middle 75% of women's diastolic blood pressures fall between 54.051 mm of Hg and 80.799 mm of Hg.



Section 1G Odd Answers

1.

a) A skewed right shape has the center on the far left and a long tail to the right. A histogram would have the highest bars on the far left with a short left tail and a long right tail.

b) A skewed left shape has the center on the far right and a long tail to the left. A histogram would have the highest bars on the far right with a short right tail and a long left tail.

c) The median average is the average or center when the data values are put in order. When data sets are not normal, we prefer to use the median as our average. The median also splits the data so that approximately 50% of the data is above the median and 50% of the data is below the median. To calculate the median, first put the numbers in order. If there is one number in the middle, then that is the median. If there are two numbers in the middle, then the median will be half way between the two numbers in the middle.

d) The first quartile is the number that approximately 25% of the data values are less than and 75% of the data values are greater than. To calculate the first quartile, simply calculate the median of the bottom half of the data when the data values are in order.

e) The third quartile is the number that approximately 75% of the data values are less than and 25% of the data values are greater than. To calculate the third quartile, simply calculate the median of the top half of the data when the data values are in order.

f) The interquartile range is the best measure of typical spread for non-normal data. It measures the distance between the middle 50% of the data values. It can also be thought of as the maximum distance between typical values in a non-normal data set. To calculate IQR, subtract the third quartile minus the first quartile.

2.

a) If the data is not normal, we should use the median as our average or center.

b) If the data is not normal, we should use the IQR as the best measure of typical spread.

c) If the data is not normal, then typical values will fall between the 1st quartile and the 3rd quartile.

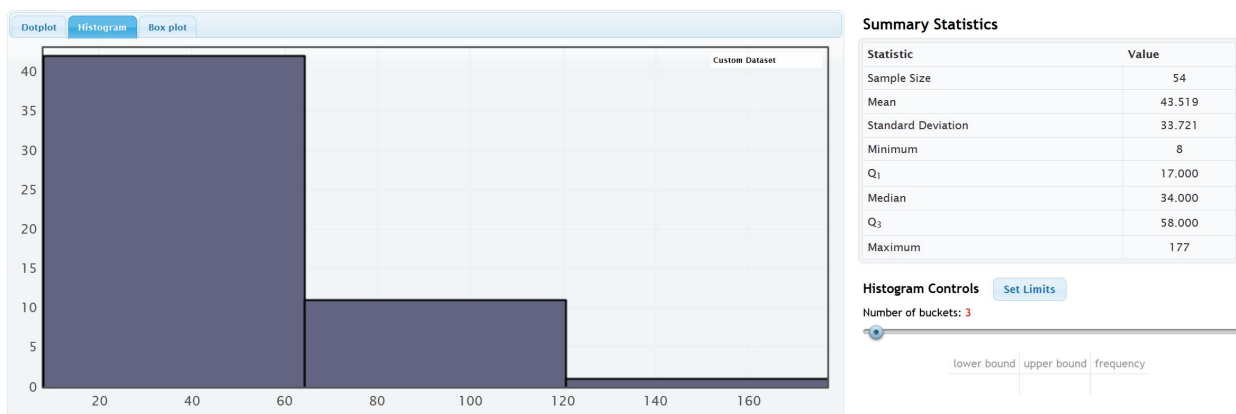
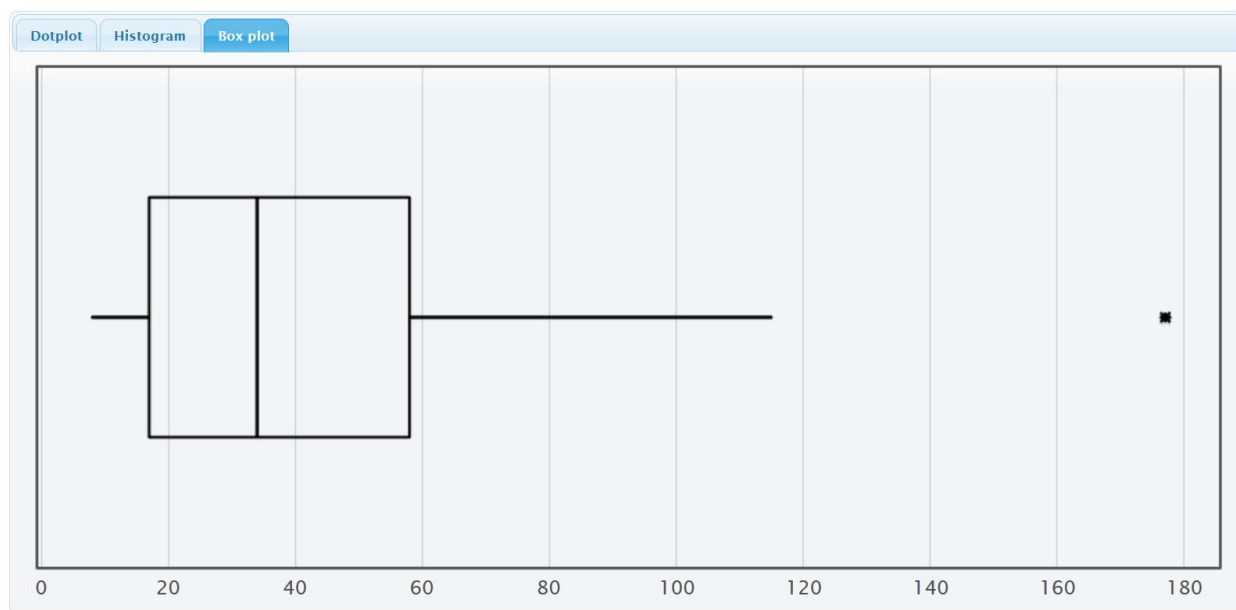
d) The middle 50% is typical for data that is not normal.



e) If the data is not normal, then you can use the box plot to identify unusually high values (high outliers). For horizontal box plots, look for circles, triangles or stars to the far right of the right whisker. For vertical box plots, look for circles, triangles or stars above the top whisker.

f) If the data is not normal, then you can use the box plot to identify unusually low values (low outliers). For horizontal box plots, look for circles, triangles or stars to the far left of the left whisker. For vertical box plots, look for circles, triangles or stars below the bottom whisker.

3.



a) The data is measuring the ages of bears. The units are months.

b) There are 54 bears in the data set. (Sample size)

c) The histogram shows that the data is skewed right.

d) The youngest bear is 8 months old.

e) The oldest bear is 177 months old.

f) Since the data is not normal, we should use the median as our average or center. The median average is 34 months, so the average age of the bears is 34 months old.



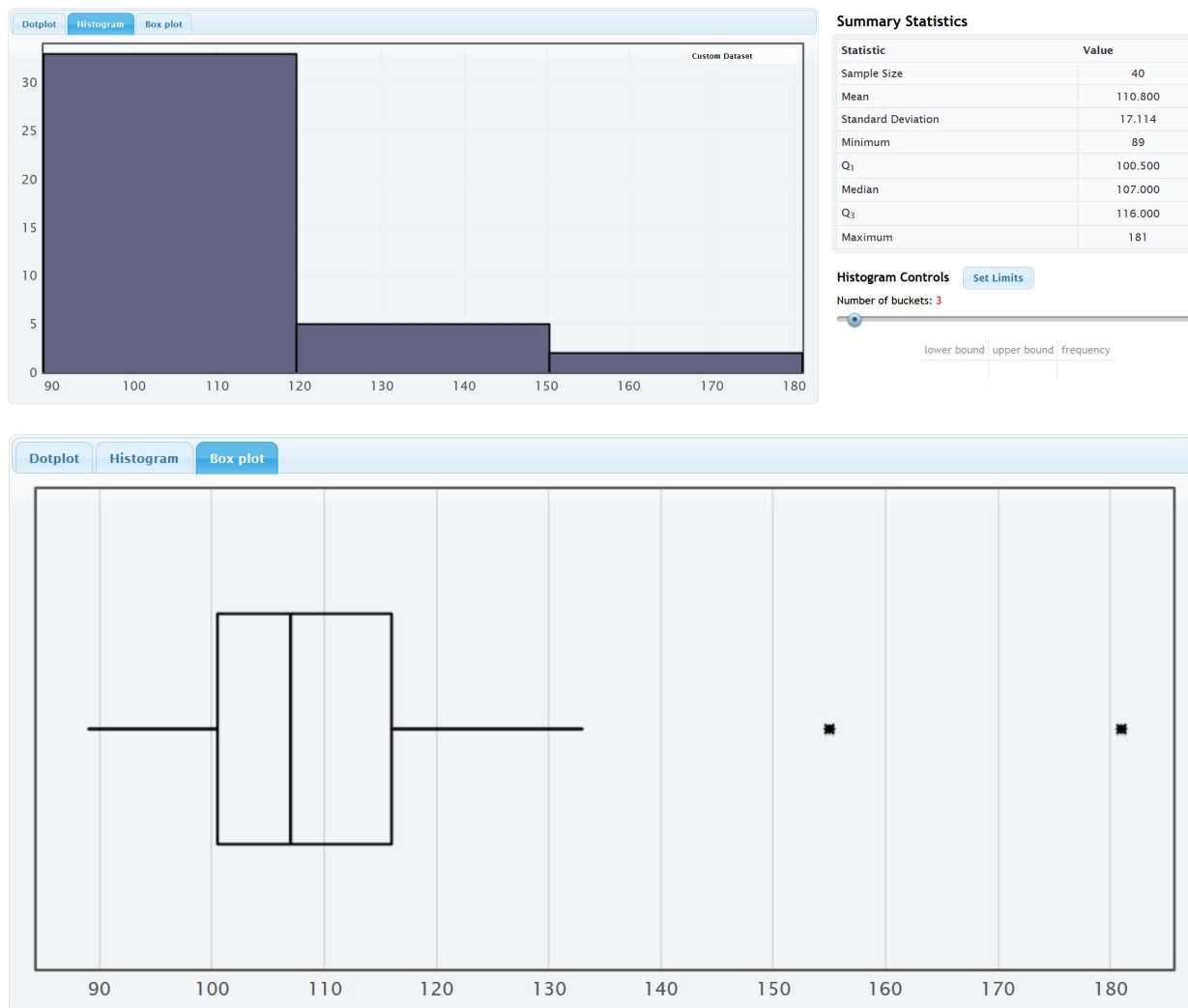
g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is the 3rd quartile minus the 1st quartile = $58 - 17 = 41$ months. So typical bear ages are within 41 months of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (17 months) and the 3rd quartile (58 months). So typical bear are between 17 months old and 58 months old.

i) The box plot has one star to the far right. This corresponds to the maximum value of 177 months. So there is only one high outlier in the data set at 177 months old.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

5.



a) The data is measuring the systolic blood pressures of women. The units are millimeters of mercury (mm of Hg).

b) There are 40 women in the data set. (Sample size)

c) The histogram shows that the data is skewed right.

d) The lowest systolic blood pressure for these women was 89 mm of Hg.

e) The highest systolic blood pressure for these women was 181 mm of Hg.



f) Since the data is not normal, we should use the median as our average or center. The median average is 107 mm of Hg, so the average systolic blood pressure for these women is 107 mm of Hg.

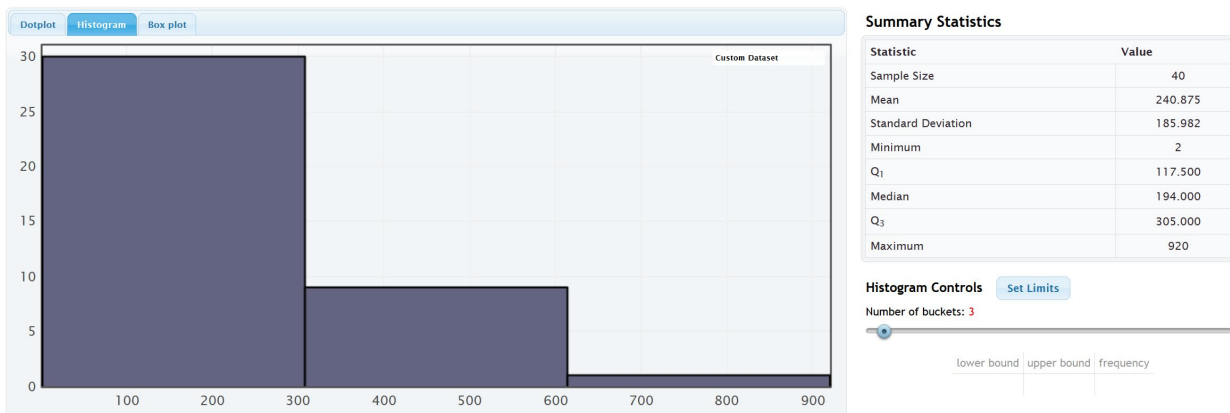
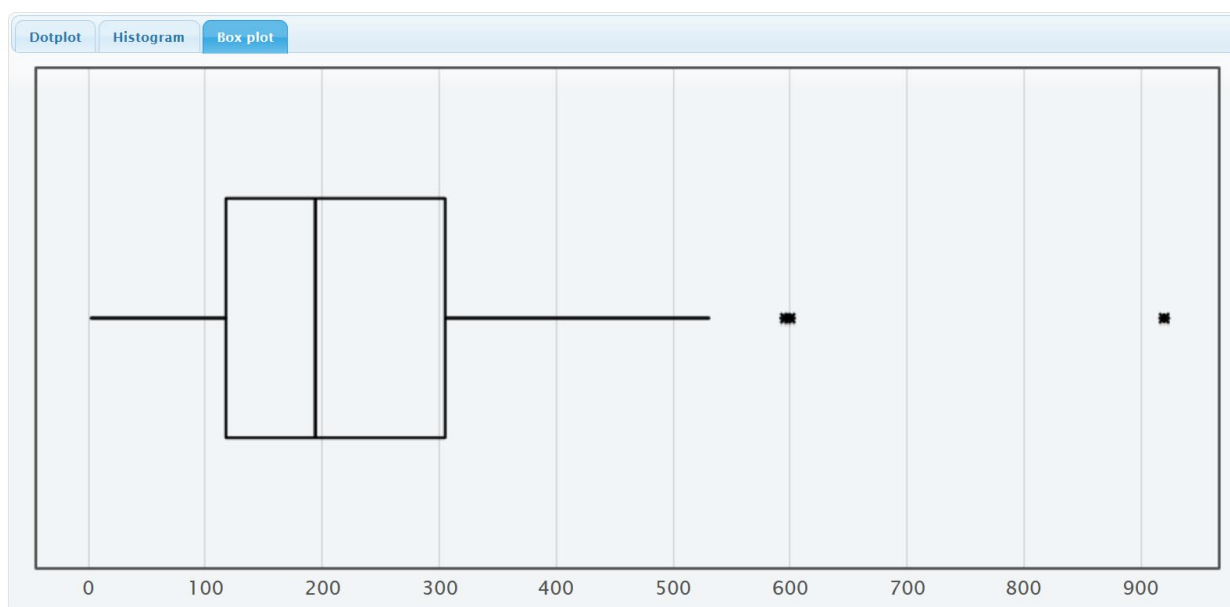
g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is the 3rd quartile minus the 1st quartile = $116 - 100.5 = 15.5$ mm of Hg. So typical women have a systolic blood pressure within 15.5 mm of Hg of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (100.5 mm of Hg) and the 3rd quartile (116 mm of Hg). So typical women in this data had a systolic blood pressure between 100.5 mm of Hg and 116 mm of Hg.

i) The box plot has two stars to the far right. This corresponds to 155 mm of Hg and the maximum value of 181 mm of Hg. So there is only two high outliers in the data set at 155 mm of Hg and 181 mm of Hg.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

7.



a) The data is measuring the cholesterol of women. The units are milligrams per deciliter (mg/dL).

b) There are 40 women in the data set. (Sample size)

c) The histogram shows that the data is skewed right.



d) The lowest cholesterol for these women was 2 mg/dL. This may be a mistake in the data. Doesn't sound possible.

e) The highest cholesterol for these women was 920 mg/dL.

f) Since the data is not normal, we should use the median as our average or center. The median average is 194 mg/dL, so the average cholesterol for these women is 194 mg/dL.

g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is the 3rd quartile minus the 1st quartile = $305 - 117.5 = 187.5$ mg/dL. So typical women have a cholesterol within 187.5 mg/dL of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (117.5 mg/dL) and the 3rd quartile (305 mg/dL). So typical women in this data had a cholesterol between 117.5 mg/dL and 305 mg/dL.

i) The box plot has three stars to the far right. So there are three unusually high cholesterol for these women. They corresponds to the cholesterol of 596 mg/dL, 600 mg/dL and 920 mg/dL.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

9.

a) The data is measuring the gas mileage for various cars. The units are in miles per gallon (mpg).

b) There are 38 cars in the data set. ("N Total")

c) The histogram shows that the data is skewed right.

d) The lowest miles per gallon was 15.5 mpg.

e) The highest miles per gallon was 37.3 mpg.

f) Since the data is not normal, we should use the median as our average or center. The median average is 24.25 mpg, so the average gas mileage for these cars is 24.25 mpg.

g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is given as 12.175 mpg. So typical cars in this data are within 12.175 mpg of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (18.425 mpg) and the 3rd quartile (30.6 mpg). So typical cars in this data have a gas mileage between 18.425 mpg and 30.6 mpg.

i) The box plot does not have any stars to the far right, so there is no high outliers in this data.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

11.

a) The data is measuring the horsepower for various cars. The units are the number of horsepower the car has.

b) There are 38 cars in the data set. ("N Total")

c) The histogram shows that the data is skewed right.

d) The smallest horsepower for these cars was 65.

e) The largest horsepower for these cars was 155.

f) Since the data is not normal, we should use the median as our average or center. The median average is 100 horsepower, so the average for these cars is 100 horsepower.

g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is given as 47.75 horsepower. So typical cars in this data are within 47.75 horsepower of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (77.25 horsepower) and the 3rd quartile (125 horsepower). So typical cars in this data have a horsepower between 77.25 and 125.



j) The box plot does not have any stars to the far right, so there is no high outliers in this data.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

13.

a) Q1 is a measure of position.

b) Mean is a measure of center.

c) Variance is a measure of spread.

d) Standard deviation is a measure of spread.

e) Minimum value is a measure of position.

f) Q3 is a measure of position.

g) The mode is a measure of center.

h) The IQR is a measure of spread.

i) The median is a measure of center.

j) The range is a measure of spread.

k) The maximum is a measure of position.

l. The midrange is a measure of center.

Chapter 1 Review Sheet All Answers

1.

a) Categorical since the data would consist of words.

b) Quantitative since it is numerical measurement data.

c) Categorical since the data would consist of words.

d) Categorical since the data would consist of words.

e) Quantitative since it is numerical measurement data.

f) Quantitative since it is numerical measurement data.

2.

a) Jim can ask every 5th student that walks into the COC cafeteria about their salary. This would have a significant amount of sampling bias.

b) Jim can put a survey on Facebook asking how money COC students make. This would have a significant amount of sampling bias.

c) Jim can have a computer randomly select student ID numbers and then track down those students whose ID numbers were selected and ask them their salary. This would have no sampling bias.

d) Jim can ask other students in his COC classes about their salary. This would have a significant amount of sampling bias since it is not a random sample.

e) Jim can randomly select 10 section numbers at COC, and then go to those classes and get data from everyone in the class. Since he chose the groups randomly, this would not have much sampling bias.

f) Jim could walk around the COC campus asking female students about their salary. Later he could walk around asking male students about their salary. Later he could compare the female and male student salaries. Since this method was not randomly selected, there would be a lot of sampling bias.



3.

Population: The collection of all people or objects to be studied. For example, a marine biologist could study all dolphins in the world.

Census: Collecting data from everyone in a population. This is the best way to collect data and minimizes sampling bias. For example, suppose our population of interest was the students at Valencia high school. We could collect data from every student at Valencia high school.

Sample: Collecting data from a small subgroup of the population. For example, if our population was all people in Palmdale, CA, we might collect data from fifty people in Palmdale.

Random: When everyone in the population has a chance to be included in the sample. Suppose our population is all COC students. We could have a computer randomly select student ID numbers and then collect data from those students.

Bias: When data does not represent the population. Asking your friends and family will not represent the population of all people in the world.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. Sample mean averages, sample standard deviations, or sample percentages would all be examples of statistics.

4.

Sampling Bias: A type of bias that results from collecting sample data that is not random or representative of the population. For example, if our population was all adults in California, and our sample consists of asking our friends and family. To limit this bias, we could take a random sample instead.

Question Bias: A type of bias that results when someone phrases the question or gives extra information with the goal of swaying the person to answer a certain way. Instead of asking a person's opinion about raising taxes, the person first gives a speech about how they think raising taxes is terrible. To limit this bias we could simply ask if the person is for raising or lowering taxes and not give any extra information.

Response Bias: A type of bias that results when people do not answer truthfully or accurately. Asking people how much they weigh in pounds will result in many people lying about the answer. Instead of asking people, we could weigh them on a scale and assure them the data will not be released.

Deliberate Bias: A type of bias that results when the people collecting the data falsify the reports, delete data, or decide to not collect data from certain groups in the population. A common deliberate bias is to delete all of the data that makes your company look bad. We could avoid this bias by not deleting data or falsifying reports. Use the data to improve the company.

Non-response Bias: A type of bias that results when people refuse to participate or give data. When calling random phone numbers to collect data, many people will refuse to answer. To limit this bias, we may leave a message asking them to call us back and offering a gift card if they do.

5. Rachael will need a group of volunteers who want to participate in the experiment. She will need to randomly assign the volunteers into two groups. One group will be the treatment group and receive actual nicotine patches. The other group will be the control group and receive a fake patch (placebo). The placebo patch and the real patch should look identical. Patches should be given to patients using a double blind approach. No volunteer in the experiment will know if they are getting the real patch or a placebo. Also those directly giving the patch will not know either. This will control the placebo effect. Randomly assigning the groups will make them alike in many confounding variables. Rachael may also exercise direct control and manipulate the groups so that they are even more alike. There are many confounding variables including the level of addiction, the number of cigarettes smoked previously, genetics, age, gender, stress, job, etc. Answers may vary. Random assignment should control these confounding variables. If the experiment shows that those with the patch have a significantly higher percentage of quitting smoking, then it will prove that using the patch causes a person to quit smoking.



6.

An experiment creates two or more similar groups with either random assignment or using the same people twice. The similar groups control confounding variables and prove cause and effect. An observational study does not create similar groups and does not control confounding variables. An observational study just collects data and analyzes it, so it cannot prove cause and effect.

Experiment Example: Suppose we want to prove that drinking alcohol causes car accidents. We can have a group of volunteers that wish to participate. We create a driving course with cones. All of the volunteers drive the course sober and we keep track of the number of cones struck. All volunteers drive the same car, with no other distractions (no phones or radio). Then we allow the volunteers to drink alcohol until they all have similar blood alcohol content. Then they can re-drive the course and we keep track of the number of cones struck. If the number of cones is significantly more in the drunk drivers, we have proven that drinking alcohol causes car accidents.

Observational Study Example: Suppose we collect data on car accidents and how many of them involved drunk driving. There are many things that influence having a car accidents other than alcohol, so this data would not prove cause and effect.

7.

- a) Identify the place value you wish to round. Look at the number to the right of the place value. If the number is 5 or above, add 1 to the place value and cut off the rest of the decimal. If the number is 4 or less, leave the place value alone and cut off the rest of the decimal.
- b) To convert a decimal proportion into a percentage, simply multiply the decimal by 100 and add on the “%” sign.
- c) To convert a percentage into a decimal proportion, remove the “%” sign, and divide the percentage by 100.
- d) To calculate a percentage divide the amount by the total.
- e) To estimate an amount, convert the percentage into a decimal proportion and multiply the proportion by the total. Round the answer to the ones place.

8.

- a) 7.22%
- b) 0.41%
- c) 56.3%
- d) 0.05%

9.

- a) 0.359
- b) 0.04823
- c) 0.00026
- d) 0.00389

10.

- a) $11/74 \approx 0.149$
- b) $0.149 \times 100\% = 14.9\%$

Approximately 14.9% of the company are managers.

- c) $27/74 \approx 0.365$
- d) $0.365 \times 100\% = 36.5\%$



Approximately 36.5% of the company are full-time employees.

e) $36/74 \approx 0.486$

f) $0.486 \times 100\% = 48.6\%$

Approximately 48.6% of the company are part-time employees.

g) Percent of Increase = $(0.365 - 0.149) / 0.149 \approx 145.0\%$ increase. This seems to be a significantly large percent of increase so the percentage of full-time employees seems significantly higher than the percentage of managers. The difference also seems to be practically significant since there are 16 more full time employees than managers and the whole company is 74 total.

h) Percent of Increase = $(0.486 - 0.365) / 0.365 \approx 33.2\%$ increase. This seems to be a large percent of increase so the percentage of part-time employees seems significantly higher than the percentage of full-time employees. The difference may not be practically significant since there are only 9 more part-time employees than full-time.

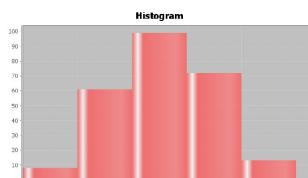
11.

$60\% = 0.6$

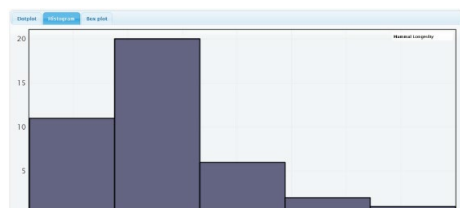
Estimated Amount = $0.6 \times 41743 \approx 25,046$ voters in Saugus.

12.

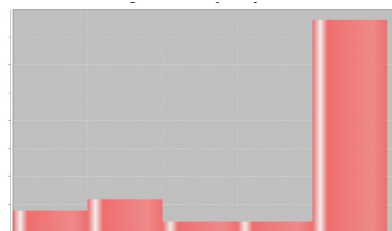
a) A normal or normally distributed histogram is unimodal and symmetric. This means that we expect the highest bar or bars to be in the middle with smaller and smaller bars as we go away from the middle. The left and right tails will be approximately the same length.



b) A skewed right or positively skewed histogram will have the highest bar or bars on the far left of the graph. It will have very few bars to the left of the center and many bars to the right of the center. Therefore the right tail will look much longer than the left tail.



c) A skewed left or negatively skewed histogram will have the highest bar or bars on the far right of the graph. It will have very few bars to the right of the center and many bars to the left of the center. Therefore, the left tail will look much longer than the right tail.



13.

- a) The first quartile (Q_1) is a measure of position. It is used to analyze typical values when data is skewed or not normal.
- b) The mean is a measure of center. It is the primary center or average when the data is normal.
- c) The variance is a measure of spread. It is used when the data is normal.
- d) The standard deviation is a measure of spread. It is the primary measure of spread when the data is normal.
- e) The minimum value is a measure of position. It can sometimes be an outlier and is used in all quantitative data regardless of shape.
- f) The third quartile (Q_3) is a measure of position. It is used to analyze typical values when data is skewed or not normal.
- g) The mode is a measure of center. It is often used in business applications or any time we wish to know the value or values that appear most often.
- h) The interquartile range (IQR) is a measure of spread. It is the primary measure of spread when the data is skewed or not normal.
- i) The median or 50th percentile or 2nd quartile (Q_2) is a measure of center. It is the primary measure of center or average when the data is skewed or not normal.
- j) The range is a measure of spread. It is usually used when someone wants a quick easy to calculate measure of spread. It does not represent typical spread.
- k) The maximum value is a measure of position. It can sometimes be an outlier and is used in all quantitative data regardless of shape.
- l) The midrange is a measure of center. It is usually used when someone wants a quick easy to calculate center or average. It may not be a very accurate average since it is often based on outliers.

14.

- a) We should use the mean average when the data is normal.
- b) We should use the median average when the data is not normal.
- c) We should use the standard deviation as our main measure of typical spread when data is normal.
- d) We should use the interquartile range (IQR) as our main measure of typical spread when data is not normal.
- e) When the data is normal, add and subtract the mean and standard deviation. Typical values will fall between $\bar{x} - s$ and $\bar{x} + s$.
- f) When data is not normal, typical values will fall between Q_1 and Q_3 .
- g) To calculate the unusually high cutoff for normal data, multiply the standard deviation by two and then add it to the mean ($\bar{x} + 2s$).
- h) To calculate the unusually high cutoff for non-normal data, multiply the IQR by 1.5 and add it to Q_3 . ($Q_3 + (1.5 \times \text{IQR})$).
- i) To calculate the unusually low cutoff for normal data, multiply the standard deviation by two and then subtract from the mean ($\bar{x} - 2s$).
- j) To calculate the unusually low cutoff for non-normal data, multiply the IQR by 1.5 and subtract it from Q_1 . ($Q_1 - (1.5 \times \text{IQR})$).



- k) To find low outliers in a normal data set, calculate the unusual low cutoff $\bar{x} - 2s$ and look for any data values that are lower than the low cutoff. To find high outliers in a normal data set, calculate the unusual high cutoff $\bar{x} + 2s$ and look for any data values that are higher than the high cutoff.
- l) To find low and high outliers in a skewed or non-normal data set, create a boxplot and look for any stars, circles or triangles outside of the whiskers.

15.

- a) The data is measuring the age of mammals. The units are in years.
- b) Sample size = 40. There are 40 mammals in the data set.
- c) The data is skewed right.
- d) The youngest mammal was 1 year old.
- e) The oldest mammal was 40 years old.
- f) Since the data was not normal, we will use the median average. The average age of the mammals is 12 years.
- g) Since the data was not normal, we will use the interquartile range to measure the typical spread. $IQR = Q3 - Q1 = 15.5 - 8 = 7.5$ years. So typical mammal ages in this data set are within 7.5 years of each other.
- h) Since the data was not normal, typical values will fall between $Q1$ and $Q3$. So typical mammal ages in this data set are between 8 years and 15.5 years.
- i) The box plot shows that there is one high outlier at 40 years.
- j) The box plot shows that there is no low outliers.

16.

- a) The data is measuring the amount of time employees have been with the company. The units are years.
- b) N total = 253. There are 253 employees in the data set.
- c) The data appears normal.
- d) The employee that has been with the company the shortest amount of time is 3.6 years.
- e) The employee that has been with the company the longest amount of time is 10.8 years.
- f) Since the data is normal, we will use the mean average. The average time that employees have been with this company is 7.345 years.
- g) Since the data is normal, we will use the standard deviation as our measure of typical spread. So typical employee times with the company are within 1.376 years of the mean.
- h) Since the data is normal, the high outlier cutoff is the mean + (2 x standard deviation) = $7.345 + (2 \times 1.376) = 10.097$. So any employees that have been with the company 10.097 years or more is considered an unusually large amount of time. The dot plot shows five employees that have been with the company from approximately 10.5 years to 10.8 years. All of these times are unusually long.
- i) Since the data is normal, the low outlier cutoff is the mean - (2 x standard deviation) = $7.345 - (2 \times 1.376) = 4.593$. So any employees that have been with the company 4.593 years or less is considered an unusually small amount of time. The dot plot shows five employees that have been with the company from approximately 3.6 years to 4.5 years. All of these times are unusually short.

17.

- a) Driving alone was most popular.
- b) Biking was least popular.
- c) 10 of the stat students walked to school.



d) 0.054

e) $0.018 \times 100\% = 1.8\%$

1.8% of the stat students used public transportation.

18.

a) 34%

b) 16%

c) Typical salaries are between 29.3 thousand dollars and 33.5 thousand dollars.

d) Salaries above 35.6 thousand dollars are considered unusually high (high outliers).

e) Salaries below 27.2 thousand dollars are considered unusually low (low outliers).

f) The average salary is 31.4 thousand dollars.

Section 2A Odd Answers

1.

N: Parameter describing the size of a population.

n: Statistic describing the size of a sample.

π or p : Parameter describing a population proportion.

\hat{p} : Statistic describing a sample proportion.

μ : Parameter describing a population mean average.

\bar{x} : Statistic describing a sample mean average.

σ : Parameter describing a population standard deviation.

s : Statistic describing a sample standard deviation.

σ^2 : Parameter describing a population variance.

s^2 : Statistic describing a sample variance.

ρ : Parameter describing a population correlation coefficient.

r : Statistic describing the correlation coefficient from sample data.

β_1 : Parameter describing a population slope.

b_1 : Statistic describing a slope from sample data.

3.

$\mu = 69.2$ inches (parameter)

$\bar{x} = 69.5$ inches (statistic)

5.

$n = 300$ students (statistic)

$\bar{x} = 101.9$ IQ (statistic)

$s = 14.8$ IQ (statistic)



7.

$\mu = 12$ units (parameter)
 $n = 160$ students (statistic)
 $\bar{x} = 12.37$ units (statistic)

9.

$N = 10,136,559$ people (parameter)

11.

$\beta_1 = 3$ pounds per month (Parameter)
 $n = 54$ bears (statistic)
 $b_1 = 2.7055$ pounds per month (Statistic)

13.

$\pi = 0.78$ (parameter)
 $n = 165$ households (statistic)
 $\hat{p} = 0.812$ (statistic)

15.

$\mu = 100$ IQ (parameter)
 $\sigma = 15$ IQ (parameter)
 $\bar{x} = 97.7$ IQ (statistic)
 $s = 15.3$ IQ (statistic)

17.

$\mu = 6.7$ pH (parameter)
 $n = 53$ lakes (statistic)
 $\bar{x} = 6.591$ pH (statistic)

19.

$N = 59,530$ people (parameter)

21.

$n = 10$ lions (statistic)
 $\bar{x} = 437.2$ pounds (statistic)
 $\mu = 420$ pounds (parameter)

23.

$b_1 = 6.23$ mpg (Statistic)
 $\beta_1 = 5.9$ mpg (Parameter)

25.

$n = 38$ cars (statistic)
 $\bar{x} = 177.289$ displacement (statistic)
 $s = 88.877$ displacement (statistic)



Sampling Distribution Act 1 Odd Answers

1.

Answers will vary. If a student got a sample mean of 131.6 cents, then the margin of error would be $131.6 - 134.338 = -2.738$ cents. The sample mean was 2.738 cents below the population mean.

3.

If all we know is one random sample, it will be virtually impossible to know the exact population mean. It is extremely difficult to determine a population parameter from a single random sample statistic.

5.

Answers will vary. If the middle 95% of sample means fell between 110 cents and 155 cents, then the approximate standard error would be $(155 - 110)/4 \approx 11.25$ cents

Sampling Distribution Act 2 Odd Answers

1.

We expect that a fair coin should have a population percentage of 50% or a population proportion of 0.5. If we flip the coin 30 times, we expect to get 15 tails.

Answers will vary. If a person got tails 19 times, then their sample proportion is approximately 0.633. That means their margin of error would be approximately $0.633 - 0.5 \approx 0.133$ or 13.3%. So our sample proportion was 0.133 above the population proportion. Or our sample percentage was 13.3% above the population percentage.

3.

If all we know is one random sample, it will be virtually impossible to know the exact population mean. It is extremely difficult to determine a population parameter from a single random sample statistic.

5.

Answers will vary. If the middle 95% of sample proportions fell between 0.33 and 0.74, then the approximate standard error would be $(0.74 - 0.33)/4 \approx 0.1025$ or 10.25%

Section 2B Odd Answers

1.

To create a sampling distribution, we take lots of random samples from a population and calculate a statistic like the sample mean average or sample proportion from each of the random samples. We then put all of the statistics on a dot plot so we can see the shape and how much variability the sample statistics have.

3.

A point estimate is when a sample statistic is used as the approximate value of a population parameter. Point estimates often create confusion in articles because the articles rarely tell you the parameter they are quoting actually came from a sample. In other words, the population parameter quoted in the article is usually just a sample statistic and therefore has a margin of error and is off from the actual population parameter.



5.

The standard error is the standard deviation of the sampling distribution. If the sampling distribution is normal, then the standard error tells us how far typical sample statistics are from the population parameter. The mean and standard deviation are only accurate when data is normal. The standard error is a type of standard deviation so is only accurate when the sampling distribution is normal.

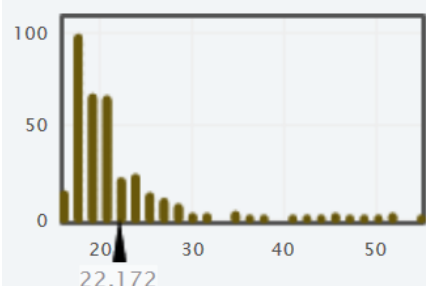
6. Do not confuse standard error with margin of error. Standard error gives us typical distance while margin of error is larger and accounts for statistics that are not typical.

7.

Population

$n = 337$, mean = 22.172

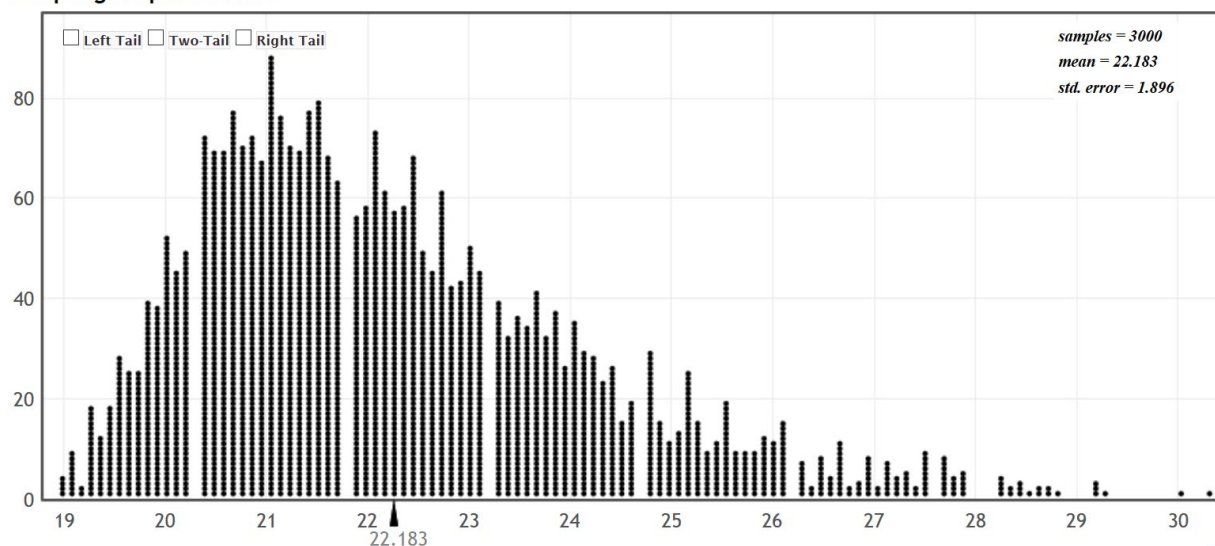
median = 20, stdev = 6.03



Custom Dataset ▾ Show Data Table Edit Data Choose samples of size $n =$ 10 Upload File Change Column(s)

Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Sampling Dotplot of Mean



a) The shape of the population is skewed right and the population mean average age of the math 140 students in fall 2015 (μ) is 22.172 years.

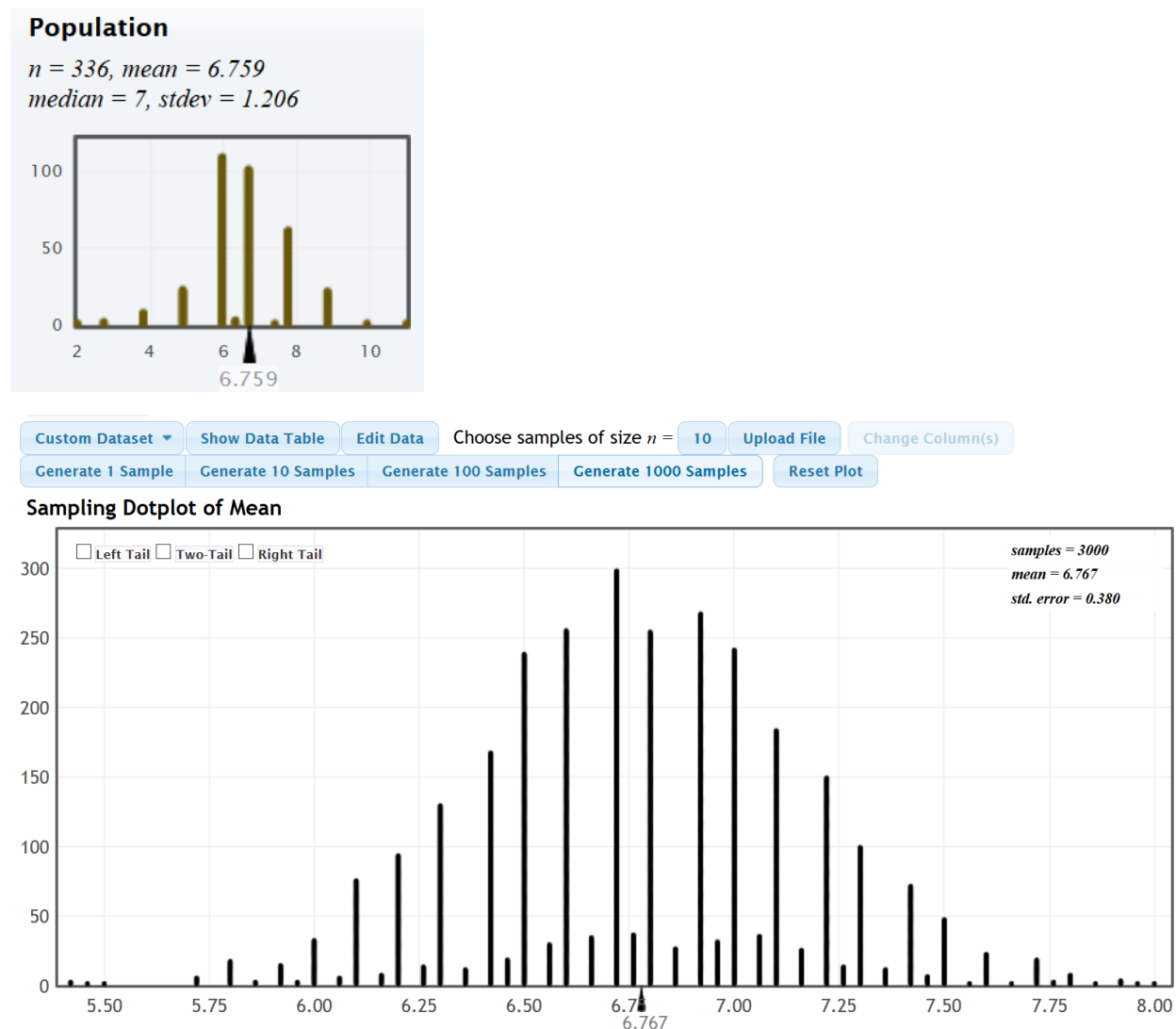
b) The samples means were generally very different than the population mean of 22.172 years. In the example sampling distribution above, the sample means fell anywhere from 18.9 years to 30.3 years. (Answers will vary.)

c) The sample means were different than each other as well. In the example sampling distribution above, the sample means fell anywhere from 18.9 years to 30.3 years. (Answers will vary.)

d) Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.



- e) The example sampling distribution above was also skewed to the right.
- f) Answers will vary. The mean average of all the sample means in the sampling distribution was 22.183 years and is pretty close to the population mean of 22.172 years.
- g) Answers will vary. The standard error for the example sampling distribution above was 1.896 years. So typical sample means were within 1.896 years from the population mean.
- 9.



- a) The shape of the population is nearly normal and the population mean average time math 140 students in fall 2015 (μ) sleep is 6.759 hours per night.
- b) The samples means were generally very different than the population mean of 6.759 hours. In the example sampling distribution above, the sample means fell anywhere from 5.4 hours to 8 hours. (Answers will vary.)
- c) The sample means were different than each other as well. In the example sampling distribution above, the sample means fell anywhere from 5.4 hours to 8 hours. (Answers will vary.)
- d) Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.
- e) The example sampling distribution above was also nearly normal.



f) Answers will vary. The mean average of all the sample means in the sampling distribution was 6.767 hours and is pretty close to the population mean of 6.759 hours.

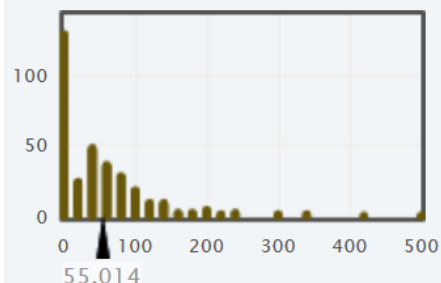
g) Answers will vary. The standard error for the example sampling distribution above was 0.380 hours. So typical sample means were within 0.380 hours from the population mean.

11.

Population

$n = 331$, mean = 55.014

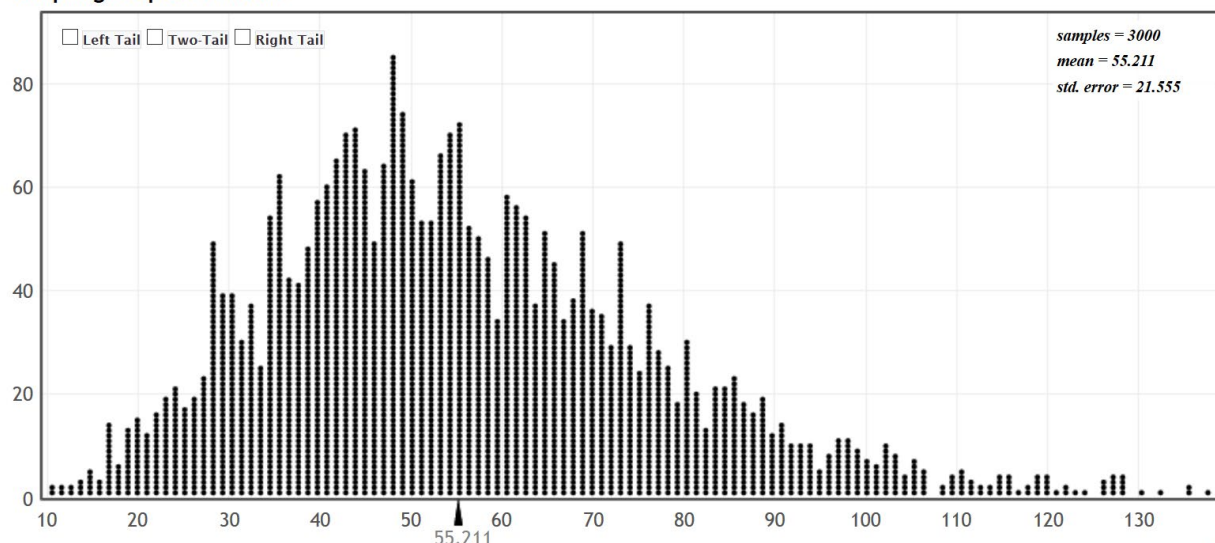
median = 45, stdev = 70.171



Custom Dataset Show Data Table Edit Data Choose samples of size $n =$ 10 Upload File Change Column(s)

Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Sampling Dotplot of Mean



a) The shape of the population is skewed right. The population mean average monthly cell phone bill for math 140 students in fall 2015 (μ) is \$55.211.

b) The samples means were generally very different than the population mean of \$55.211. In the example sampling distribution above, the sample means fell anywhere from \$9.60 to \$137.50. (Answers will vary.)

c) The sample means were different than each other as well. In the example sampling distribution above, the sample means fell anywhere from \$9.60 to \$137.50. (Answers will vary.)

d) Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.

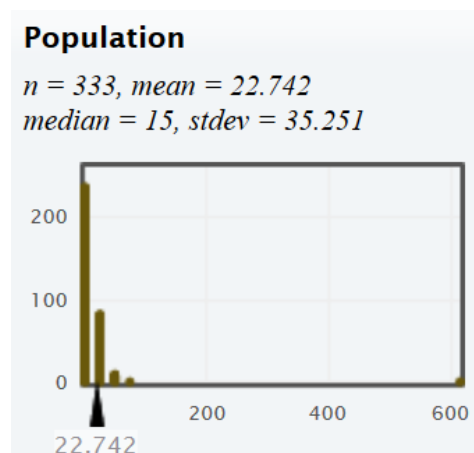
e) The example sampling distribution above was also skewed to the right.



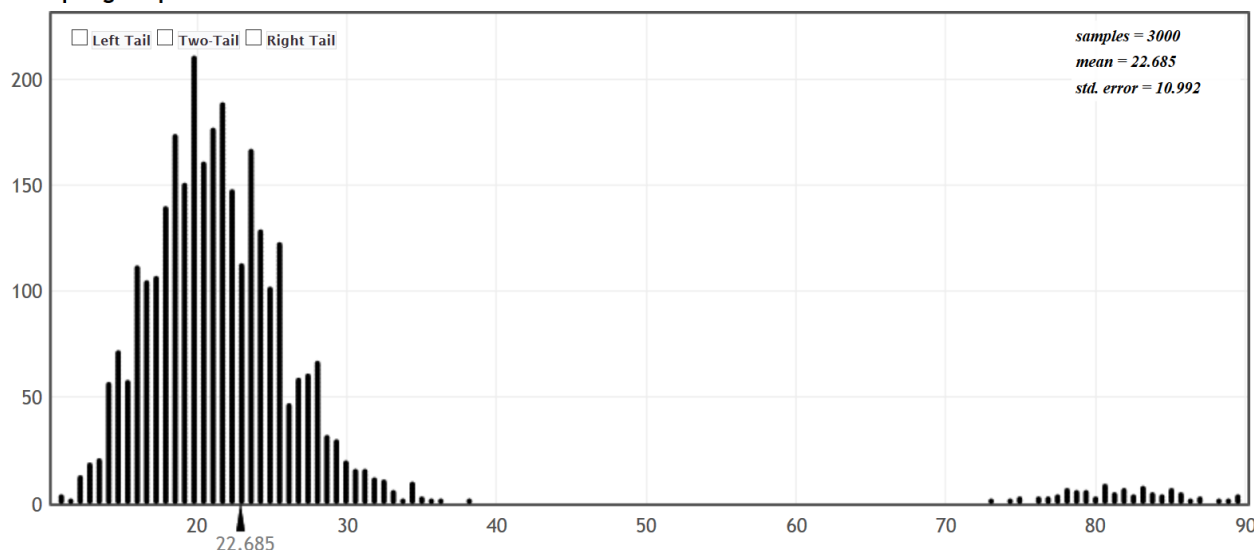
f) Answers will vary. The mean average of all the sample means in the sampling distribution was \$55.211 and is the same as the population mean of \$55.211.

g) Answers will vary. The standard error for the example sampling distribution above was \$21.555. So typical sample means were within \$21.555 from the population mean.

13.



Sampling Dotplot of Mean



- The shape of the population is skewed right. The population mean average time math 140 students in fall 2015 take to get to school (μ) is 22.742 minutes.
- The samples means were generally very different than the population mean of 22.742 minutes. In the example sampling distribution above, the sample means fell anywhere from 10.4 minutes to 89.5 minutes. (Answers will vary.)
- The sample means were different than each other as well. In the example sampling distribution above, the sample means fell anywhere from 10.4 minutes to 89.5 minutes. (Answers will vary.)
- Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.
- The example sampling distribution above was also skewed to the right.
- Answers will vary. The mean average of all the sample means in the sampling distribution was 22.685 minutes and is very close to the population mean of 22.742 minutes.

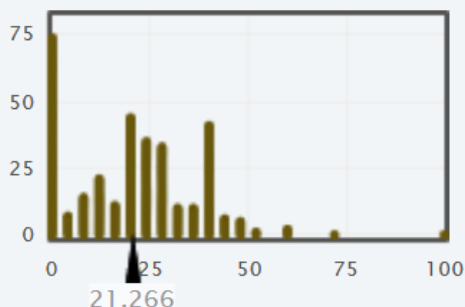


g) Answers will vary. The standard error for the example sampling distribution above was 10.992 minutes. So typical sample means were within 10.992 minutes from the population mean.

15.

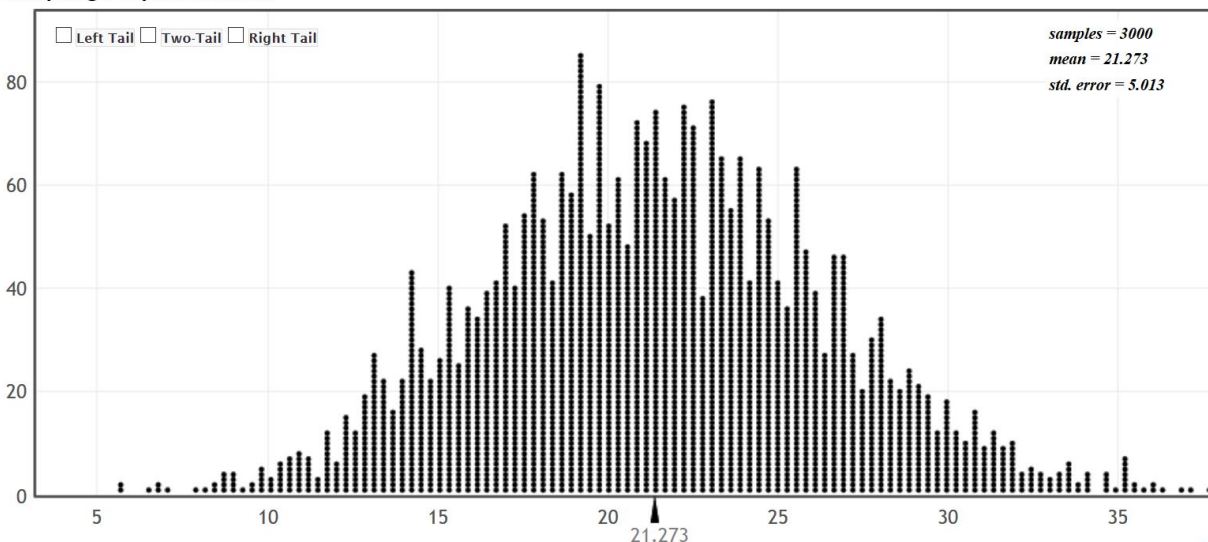
Population

$n = 331$, $mean = 21.266$
 $median = 20$, $stdev = 16.104$



Custom Dataset ▾ Show Data Table Edit Data Choose samples of size $n =$ 10 Upload File Change Column(s)
 Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Sampling Dotplot of Mean



a) The shape of the population is skewed right. The population mean average time per week that math 140 students in fall 2015 work (μ) is 21.266 hours.

b) The sample means were generally very different than the population mean of 21.266 hours. In the example sampling distribution above, the sample means fell anywhere from 5.6 hours to 37.6 hours. (Answers will vary.)

c) The sample means were different than each other as well. In the example sampling distribution above, the sample means fell anywhere from 5.6 hours to 37.6 hours. (Answers will vary.)

d) Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.

e) The example sampling distribution above was nearly normal.

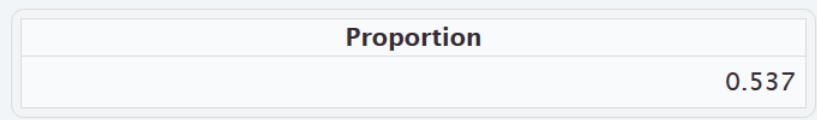
f) Answers will vary. The mean average of all the sample means in the sampling distribution was 21.273 hours and is very close to the population mean of 21.266 hours.



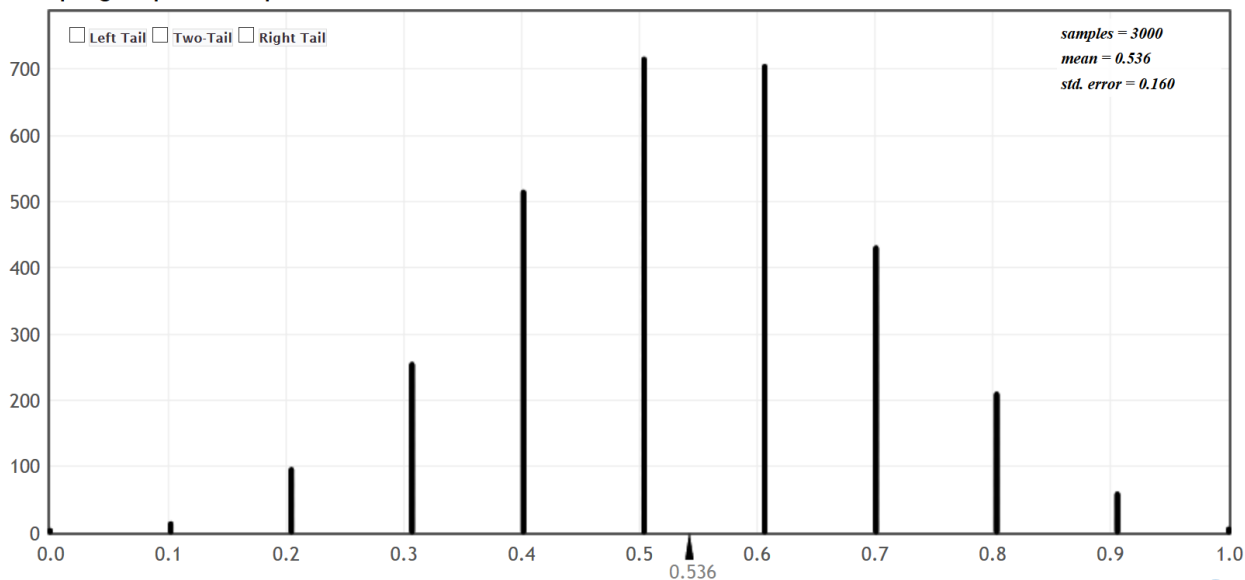
g) Answers will vary. The standard error for the example sampling distribution above was 5.013 hours. So typical sample means were within 5.013 hours from the population mean.

17.

Original Population



Sampling Dotplot of Proportion



- The sample proportions were generally very different than the population proportion of 0.537 (53.7%). In the example sampling distribution above, the sample proportions fell anywhere from 0.1 (10%) to 1.0 (100%). (Answers will vary.)
- The sample proportions were different than each other as well. In the example sampling distribution above, the sample proportions fell anywhere from 0.1 (10%) to 1.0 (100%). (Answers will vary.)
- Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.
- The example sampling distribution above was nearly normal.
- Answers will vary. The mean average of all the sample proportions in the sampling distribution was 0.536 and is very close to the population proportion of 0.537.
- Answers will vary. The standard error for the example sampling distribution above was 0.160. So typical sample proportions were within 0.160 (16%) from the population proportion.



19.

Original Population

Proportion
0.091

Custom Data ▾

Edit Proportion

Edit Data

Choose samples of size $n = 10$

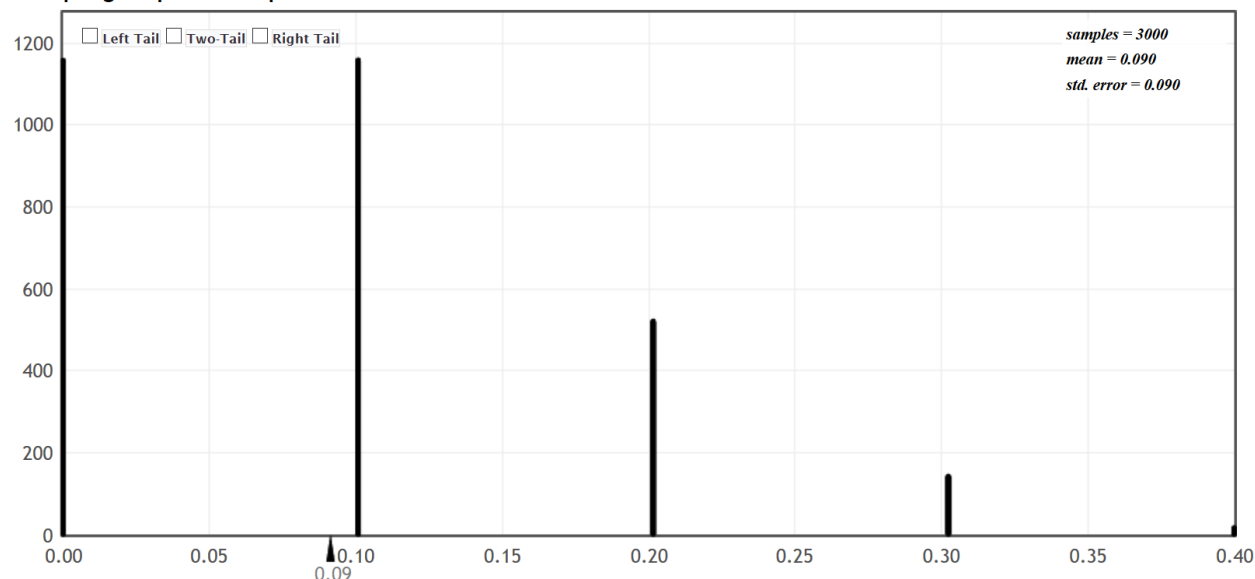
Generate 1 Sample

Generate 10 Samples

Generate 100 Samples

Generate 1000 Samples

Reset Plot

Sampling Dotplot of Proportion

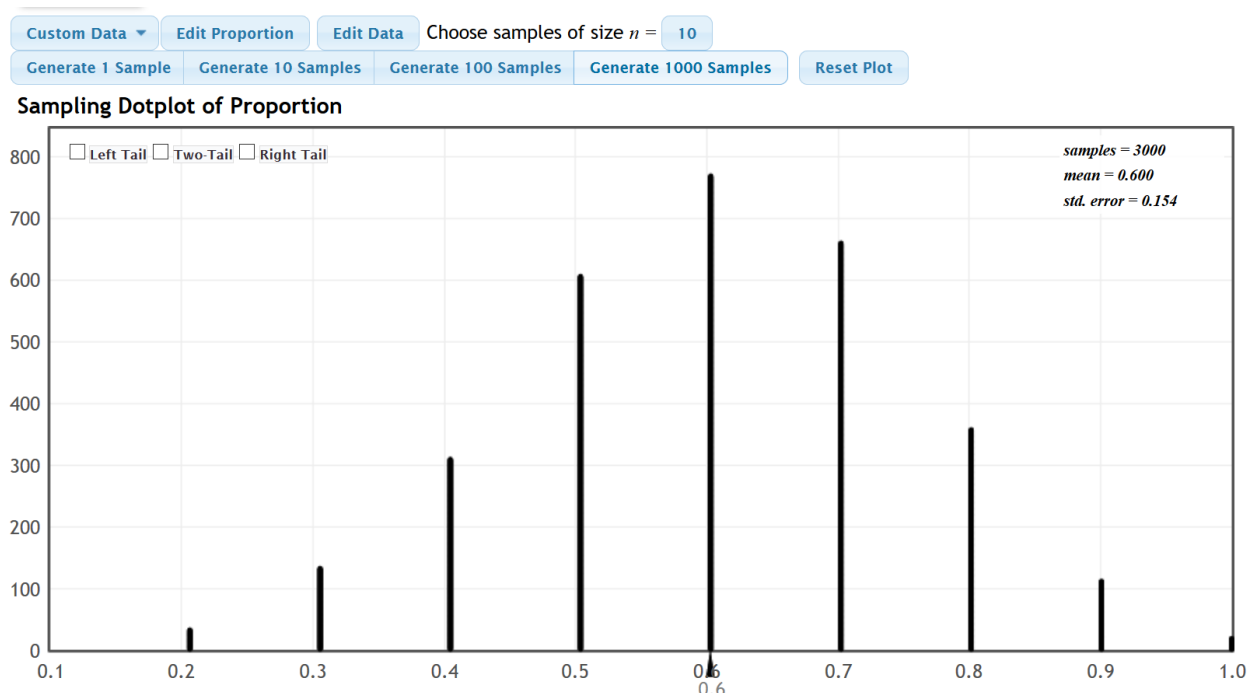
- a) The sample proportions were generally very different than the population proportion of 0.091 (9.1%). In the example sampling distribution above, the sample proportions fell anywhere from 0 (0%) to 0.4 (40%). (Answers will vary.)
- b) The sample proportions were different than each other as well. In the example sampling distribution above, the sample proportions fell anywhere from 0 (0%) to 0.4 (40%). (Answers will vary.)
- c) Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.
- d) The example sampling distribution above used samples of size $n = 10$ and was skewed to the right.
- e) Answers will vary. The mean average of all the sample proportions in the sampling distribution was 0.09 and is very close to the population proportion of 0.091.
- f) Answers will vary. The standard error for the example sampling distribution above was 0.09. So typical sample proportions were within 0.09 (9%) from the population proportion.

21.

Original Population

Proportion
0.6





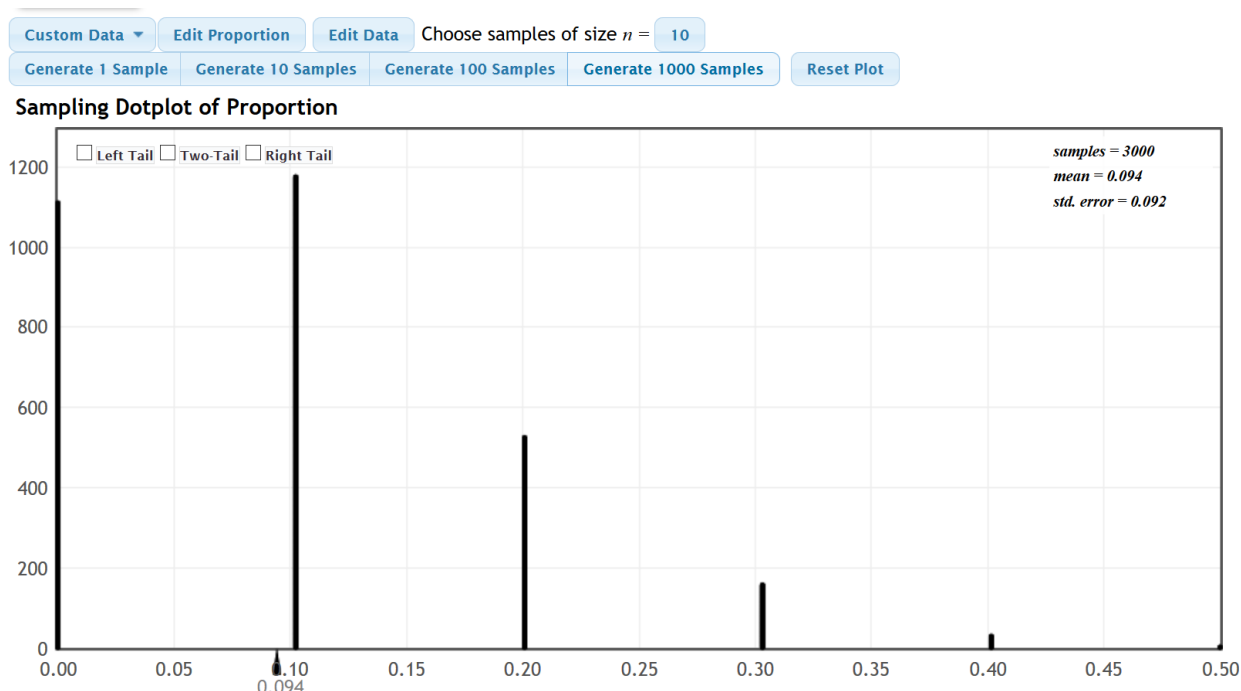
- a) The sample proportions were generally very different than the population proportion of 0.6 (60%). In the example sampling distribution above, the sample proportions fell anywhere from 0.2 (20%) to 1.0 (100%). (Answers will vary.)
- b) The sample proportions were different than each other as well. In the example sampling distribution above, the sample proportions fell anywhere from 0.2 (20%) to 1.0 (100%). (Answers will vary.)
- c) Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.
- d) The example sampling distribution above used samples of size $n = 10$ and was nearly normal.
- e) Answers will vary. The mean average of all the sample proportions in the sampling distribution was 0.6 and is the same as the population proportion of 0.6.
- f) Answers will vary. The standard error for the example sampling distribution above was 0.154. So typical sample proportions were within 0.154 (15.4%) from the population proportion.

23.

Original Population

Proportion
0.094





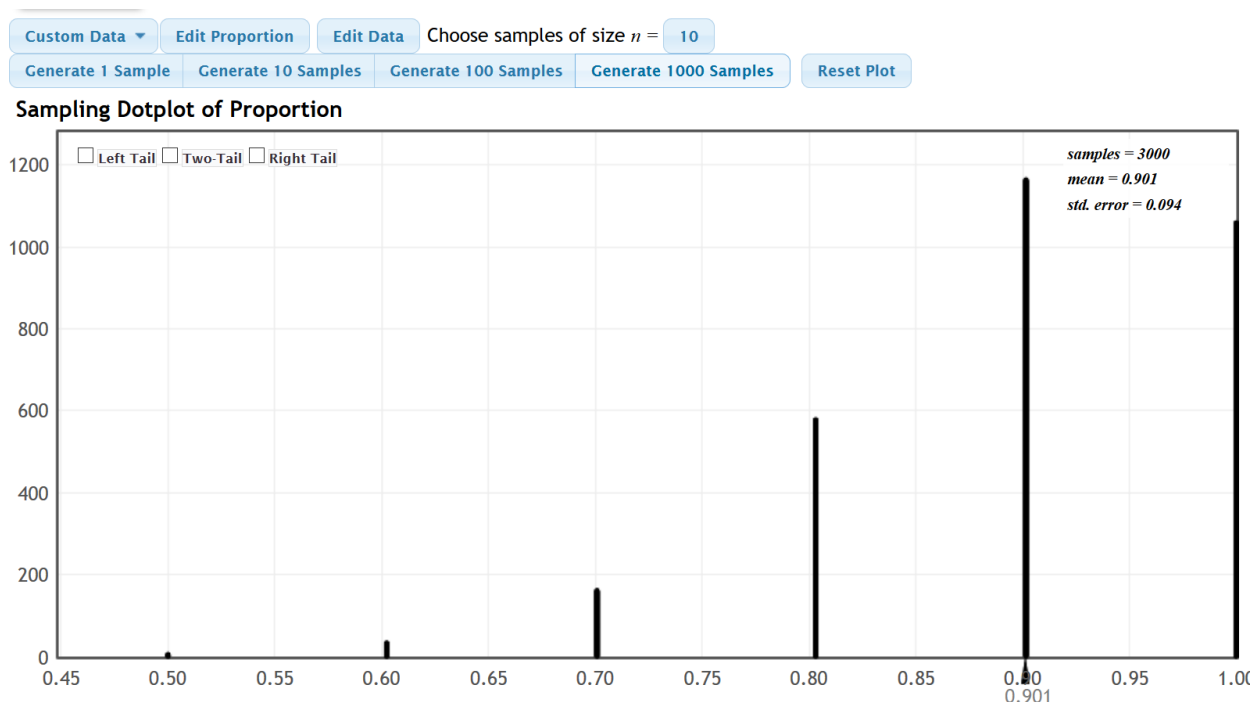
- a) The sample proportions were generally very different than the population proportion of 0.094 (9.4%). In the example sampling distribution above, the sample proportions fell anywhere from 0 (0%) to 0.4 (40%). (Answers will vary.)
- b) The sample proportions were different than each other as well. In the example sampling distribution above, the sample proportions fell anywhere from 0 (0%) to 0.4 (40%). (Answers will vary.)
- c) Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.
- d) The example sampling distribution above used samples of size $n = 10$ and was skewed right.
- e) Answers will vary. The mean average of all the sample proportions in the example sampling distribution above was 0.094 and is the same as the population proportion of 0.094.
- f) Answers will vary. The standard error for the example sampling distribution above was 0.092. So typical sample proportions were within 0.092 (9.2%) from the population proportion.

25.

Original Population

Proportion
0.9





- The sample proportions were generally very different than the population proportion of 0.9 (90%). In the example sampling distribution above, the sample proportions fell anywhere from 0.5 (50%) to 1.0 (100%). (Answers will vary.)
- The sample proportions were different than each other as well. In the example sampling distribution above, the sample proportions fell anywhere from 0.5 (50%) to 1.0 (100%). (Answers will vary.)
- Answers will vary. The example sampling distribution above took 3000 samples all of size $n = 10$.
- The example sampling distribution above used samples of size $n = 10$ and was skewed left.
- Answers will vary. The mean average of all the sample proportions in the example sampling distribution above was 0.901 and is very close to the population proportion of 0.90.
- Answers will vary. The standard error for the example sampling distribution above was 0.094. So typical sample proportions were within 0.094 (9.4%) from the population proportion.

Section 2C Odd Answers

1.

Normality in sampling distributions are important for several reasons. We often use the mean average of all of the sample statistics in the distribution as an approximation of the population parameter. Mean averages are only accurate when they come from data that is normally distributed. We also use the standard deviation of the sampling distribution (standard error) as a measure of how far typical statistics are from the population parameter. Standard Deviation is also only accurate if it comes from data that is normally distributed. There are a multitude of formulas in statistics that use standard error to understand populations. That is why formulas are often tied to requirements that the sampling distribution be normal.

3.

The central limit theorem states that if the sample size is sufficiently large, then a sampling distribution for sample means or sample proportions will be normally distributed. The idea is to explore what conditions will indicate that our sampling distribution will be normal. Remember, normality is important when using the mean and standard deviation



of the sampling distribution. For sample means from a non-normal population, we like our sample size to be at least 30. If a population was already normal, then we can have sample sizes below 30 for sample means. For sample proportions, we need to have at least ten successes and at least ten failures in our categorical data for the sampling distribution to be normal.

5.

Less data, results in more variability. So if we decrease the sample size and the population is not normal, then the standard error will get larger and the sampling distribution will look less and less normal.

7.

Yes. If the population was already normal, then the sampling distribution for sample means will look even more normal than the population. This happens because the standard error is smaller than the population standard deviation.

9.

a) The shape of the population is skewed right and the population mean average is 22.172 years.

b) This population is not normal. If we collected a random sample from this population, we must have a sample size of 30 or more.

11.

a) The shape of the population is skewed right and the population mean average is 55.014 dollars per month.

b) This population is not normal. If we collected a random sample from this population, we must have a sample size of 30 or more.

13.

a) The shape of the population is nearly normal and the population mean average is 66.511 inches.

b) This population is normal so any sampling distribution taken from this data will be even more normal. If we collected a random sample from this population, any sample size will give a normal sampling distribution. Though of course, more random data is better.

15.

a) $n = 10/0.091 \approx 109.89$

Always round sample size requirements up. So if we have a sample size of 110 or more, we are likely to have at least 10 successes.

b) $n = 10/(1 - 0.091) \approx 11.001$

Always round sample size requirements up. So if we have a sample size of 12 or more, we are likely to have at least 10 failures.

c) We will take the larger of the success and failure sample size requirements. We should recommend collecting data with a sample size of 110 or higher. In that case, we can expect the sampling distribution for sample proportions to be nearly normal.

17. $9.4\% = 0.094$

a) $n = 10/0.094 \approx 106.38$

Always round sample size requirements up. So if we have a sample size of 107 or more, we are likely to have at least 10 successes.



b) $n = 10/(1 - 0.094) \approx 11.038$

Always round sample size requirements up. So if we have a sample size of 12 or more, we are likely to have at least 10 failures.

c) We will take the larger of the success and failure sample size requirements. We should recommend collecting data with a sample size of 107 or higher. In that case, we can expect the sampling distribution for sample proportions to be nearly normal.

19. $10\% = 0.1$

a) $n = 10/0.1 = 100$

So if we have a sample size of 100 or more, we are likely to have at least 10 successes.

b) $n = 10/(1 - 0.1) \approx 11.111$

Always round sample size requirements up. So if we have a sample size of 12 or more, we are likely to have at least 10 failures.

c) We will take the larger of the success and failure sample size requirements. We should recommend collecting data with a sample size of 100 or higher. In that case, we can expect the sampling distribution for sample proportions to be nearly normal.

Section 2D Odd Answers

1.

$4.9\% \pm 1.3\%$

Interval Notation: $(3.6\%, 6.2\%)$ or $(0.036, 0.062)$

Inequality Notation: $0.036 < \pi < 0.062$

Sentence: We are 95% confident that the population percentage of adults with this disease is between 3.6% and 6.2%.

3.

$17.11 \text{ mm of Hg} \pm 3.31 \text{ mm of Hg}$

Interval Notation: $(13.80 \text{ mm of Hg}, 20.42 \text{ mm of Hg})$

Inequality Notation: $13.80 \text{ mm of Hg} < \sigma < 20.42 \text{ mm of Hg}$

Sentence: We are 90% confident that the population standard deviation for women's systolic blood pressure is between 13.80 mm of Hg and 20.42 mm of Hg.

5.

$15.98 \text{ thousand dollars} \pm 3.78 \text{ thousand dollars}$

Interval Notation: $(12.20 \text{ thousand dollars}, 19.76 \text{ thousand dollars})$

Inequality Notation: $12.20 \text{ thousand dollars} < \mu < 19.76 \text{ thousand dollars}$

Sentence: We are 90% confident that the population mean average price of a used mustang car is between 12.20 thousand dollars and 19.76 thousand dollars.



7.

172.55 pounds \pm 11.272 pounds

Interval Notation: (161.278 pounds , 183.822 pounds)

Inequality Notation: 161.278 pounds $< \mu <$ 183.822 pounds

Sentence: We are 99% confident that the population mean average weight of men is between 161.278 pounds and 183.822 pounds.

9.

36.9% \pm 1.44%

Interval Notation: (35.46% , 38.34%) or (0.3546 , 0.3834)

Inequality Notation: 0.3546 $< \pi <$ 0.3834

Sentence: We are 95% confident that the population percentage of women in the U.S that are overweight is between 35.46% and 38.34%.

11.

Sentence: We are 95% confident that the population proportion of fat in the milk from Jersey cows is between 4.6% and 5.2%.

Sample proportion $\hat{p} = (0.052 + 0.046)/2 = 0.049$ Margin of Error = $(0.052 - 0.046)/2 = 3 \times 10^{-3} = 0.003$

13.

Sentence: We are 90% confident that the population proportion of people that will vote for the independent party candidate is between 6.8% and 8.3%.

Sample proportion $\hat{p} = (0.083 + 0.068)/2 = 0.0755$ Margin of Error = $(0.083 - 0.068)/2 = 7.5 \times 10^{-3} = 0.0075$

15.

Sentence: We are 99% confident that the population standard deviation for the heights of men is between 2.34 inches and 2.87 inches.

Sample standard deviation $s = (2.87 + 2.34)/2 = 2.605$ inchesMargin of Error = $(2.87 - 2.34)/2 = 0.265$ inches

17.

We are 99% confident that the population mean average pH of lakes in Florida is between 6.118 and 7.064.

Sample mean $\bar{x} = (7.064 + 6.118)/2 = 6.591$ Margin of Error = $(7.064 - 6.118)/2 = 0.473$

19.

We are 95% confident that the population mean average price of apartments in Manhattan, NY is between \$2514.36 and \$3798.64.

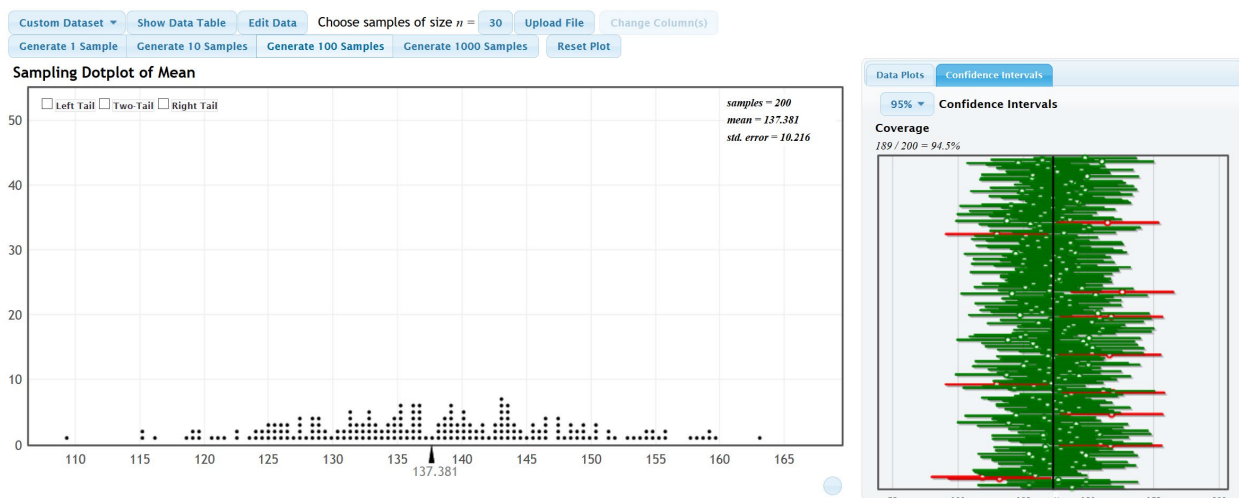
Sample mean $\bar{x} = (3798.64 + 2514.36)/2 = \3156.50 Margin of Error = $(3798.64 - 2514.36)/2 = \642.14 

21.

By making so many random samples, we see sampling variability at work. Each random sample is different. Each random sample has a different mean and different individual values. This results in a different confidence interval. Some random sample means are so vastly different from the population mean, that the confidence interval created was red and did not even contain the population mean. Other random sample means that were closer to the population mean and the confidence intervals created from these samples were green and did contain the population mean. So sampling variability suggests that not all confidence intervals made from random samples will contain the population mean.

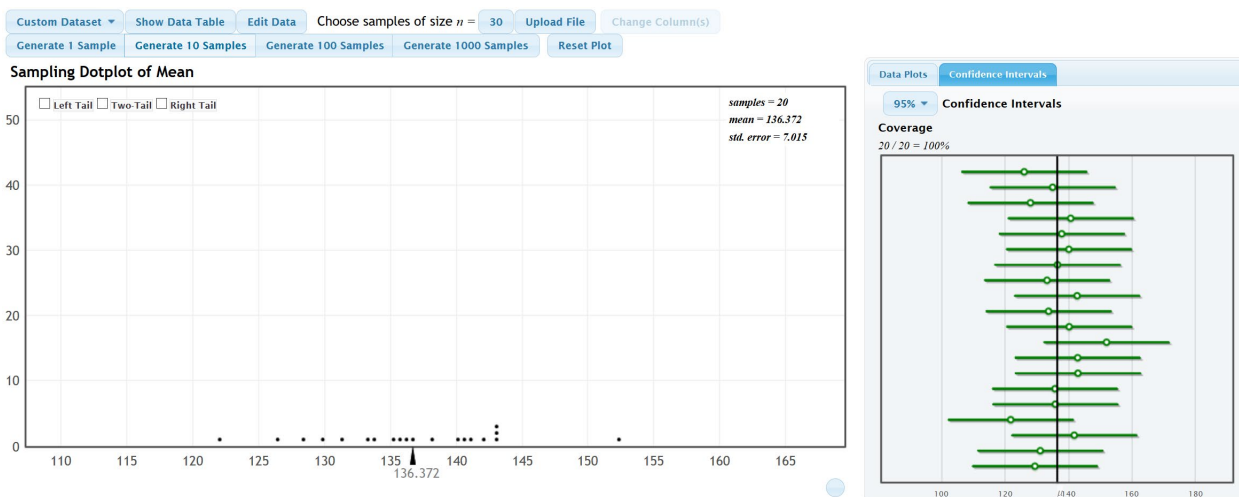
23.

Answers will vary. The example printout below, we took a total of 200 random samples and created 200 different confidence intervals. 189 of the confidence interval contained the population mean. So 94.5% of the confidence intervals contained the population mean. Notice that this is pretty close to the 95% confidence level.



25.

Answers will vary. For smaller number of samples the percentage that contain the population value is farther off of 95%. As the number of samples increased, the percentage got closer to 95%. The example printout below shows that at 20 samples, the percentage was still 100%, but the printout above shows that at 200 samples were at 94.5% which is much closer to the 95% confidence level. This is not surprising, since more random data usually results in less variability. The percentage from 200 random data sets generally will be more accurate than 20 data sets.

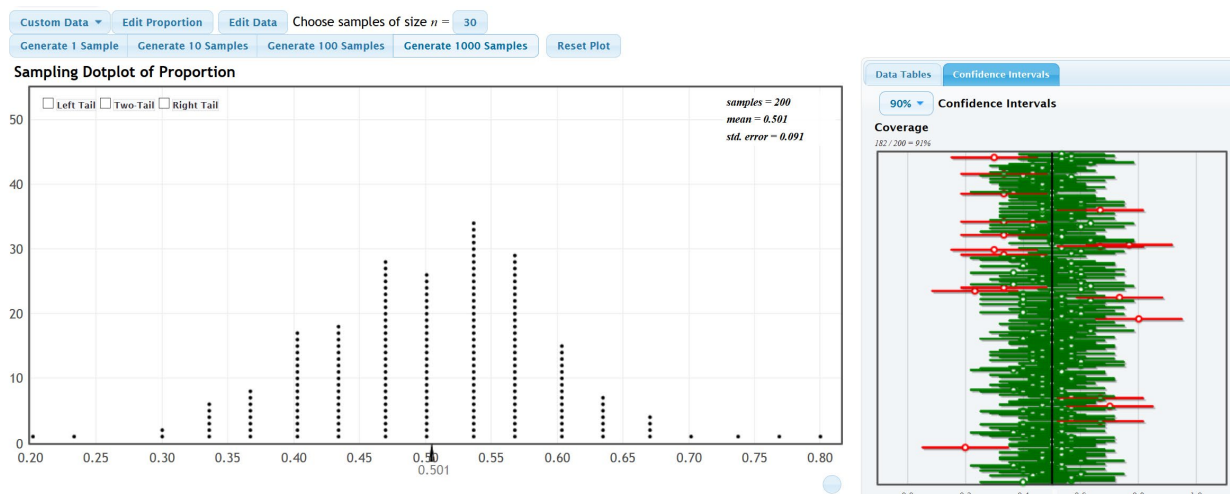


27.

By making so many random samples, we see sampling variability at work. Each random sample is different. Each random sample has a different proportion and different individual values. This results in a different confidence interval. Some random sample proportions are so vastly different than the population proportion, that the confidence interval created was red and did not contain the population proportion. Other random sample proportions were closer to the population proportion and the confidence intervals created from these samples were green and did contain the population proportion. So sampling variability suggests that not all confidence intervals made from random samples will contain the population proportion.

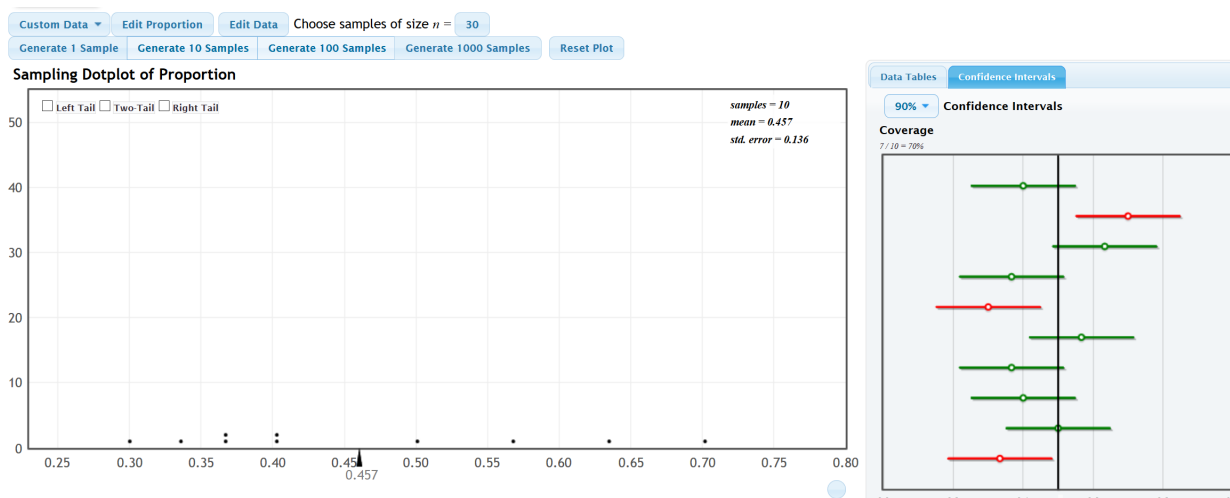
29.

Answers will vary. The example printout below, we took a total of 200 random samples and created 200 different confidence intervals. 182 of the confidence interval contained the population mean. So 91% of the confidence intervals contained the population proportion. Notice that this is pretty close to the 90% confidence level.



31.

Answers will vary. For smaller number of samples the percentage that contain the population proportion is farther off of 90%. As the number of samples increased, the percentage got closer to 90%. The example printout below shows that at 10 samples, the percentage was 70%, but the printout above shows that at 200 samples was 91% which is much closer to the 90% confidence level. This is not surprising, since more random data usually results in less variability. The percentage from 200 random data sets generally will be more accurate than 10 data sets.



Section 2E Odd Answers

1.

One-population Proportion Assumptions

- The categorical sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least ten successes and at least ten failures.

3.

One-population Bootstrap Assumptions

- The sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.

5.

a)

The sample data did not pass all of the assumptions.

Random Sample? Yes. Assuming it is a random sample.

Individual rats independent? Yes. A small random sample of rats out of a large population will likely be independent of each other.

At least ten success? Yes. There were 23 rats that showed empathy.

At least ten failures? No. There were only 7 rats that did not show empathy.

b)

We estimate that the sample proportion can be 0.199 off from the population proportion.

c)

We are 99% confident that the population proportion of rats that show empathy is between 0.5678 and 0.9656.

OR

We are 99% confident that the population percentage of rats that show empathy is between 56.78% and 96.56%.

7.

a)

The sample data did pass all of the assumptions.

Random Sample? Yes. Given in #6 that it is a random sample.

Individual tests independent? Yes. A small random sample of lie detector tests out of a large population will likely be independent of each other. We would not want all of the tests done on the same machine though.

At least ten success? Yes. The machine caught the lie 31 times.

At least ten failures? Yes. The machines did not catch the lie 17 times.

b)

We estimate that the sample proportion can be 0.135 off from the population proportion.



c)

We are 95% confident that the population proportion of lies caught by lie detector machines is between 0.5105 and 0.7811.

OR

We are 95% confident that the population percentage of lies caught by lie detector machines is between 51.05% and 78.11%.

9.

a)

This data does not meet the assumptions. The confidence interval may not represent the population very well.

Random Sample? Yes. Given in the problem.

Individuals independent? Maybe not. The population size may be rather small compared to the sample. Individual cereals may be related.

At least 10 successes? No. There were only 4 cereals made by Quaker.

At least 10 failures? Yes. There were 20 cereals in the sample data not made by quaker.

b)

The population proportion could be as much as 0.125 from the sample proportion.

c)

We are 90% confident that the population proportion of cereals made by Quaker is in between 0.0415 and 0.2918.

OR

We are 90% confident that the population percentage of cereals made by Quaker is in between 4.15% and 29.18%.

Since the sample data did not meet the assumptions, the confidence interval may not be accurate.

11.

a)

Random Sample? Yes. Given in the problem.

Individuals independent? Probably not. The population size is rather small and certain companies may tend to put more or less sugar in their cereals.

At least ten successes? Yes. There were 10 cereals with high sugar content.

At least ten failures? Yes. There were 14 cereals that did not have high sugar content.

If the cereals are independent, this would pass the assumptions.

b)

The sample proportion of cereals with high sugar can be 0.197 off from the population proportion.



c)

We are 95% confident that the population proportion of cereals with high sugar content is between 0.2194 and 0.6139.

OR

We are 95% confident that the population percentage of cereals with a high sugar content is between 21.94% and 61.39%.

13.

a)

Random sample? Yes. Given in #12.

Individuals Independent? Probably. The population of all students that take ACT is rather large. Individuals in a small random sample will likely be independent.

At least 30 or normal? Yes. The data was skewed left, but because the sample size is 45 and over 30 it will pass the "30 or normal" requirement.

b)

The sample mean ACT score of 20.8 could be as much as 2.472 points off from the population mean.

c)

We are 90% confident that the population mean average ACT score is between 18.3284 points and 23.2716 points.

15.

a)

Random Sample? Yes. Given in #14.

Individuals Independent? Probably. The population is very large so a small sample of 50 randomly selected people will likely be independent of each other.

At least 30 or normal? Yes. The histogram looks normal (bell shaped). Also, the sample size was 50 which is over 30.

b)

The sample mean average body temperature of 98.26 °F could be as much as 0.217 °F off from the population mean.

c)

We are 95% confident that the population mean average body temperature is between 98.0426 °F and 98.4774 °F.

17.

a)

The data does not pass the assumptions for a mean average confidence interval. The confidence interval may not represent the population mean very well.

Random Sample? Yes. Given in #16.

Individuals Independent? Probably not. The population size is rather small and certain companies may put more or less sugar in their cereals.

At least 30 or normal? No. The sample size was only 24 and the data looks uniform (or bimodal) and is not normal.



b)

The sample mean amount of sugar in these cereals of 7.208 grams could be as much as 2.656 grams off from the population mean amount of sugar in all cereals.

c)

We are 99% confident that the population mean average amount of sugar in cereals is between 4.5527 grams and 9.8639 grams.

This confidence interval may not be accurate since the data used did not pass the assumptions.

19.

a)

If the individual cereals are independent of each other, the data would pass the assumptions for a mean average confidence interval.

Random Sample? Yes. Given in #18.

Individuals Independent? Probably not. The population size is rather small and certain companies may have more or less carbs in their cereals.

At least 30 or normal? Yes. The sample size was only 24. However, the histogram looks normal (bell shaped) so it does pass the "30 or normal" requirement.

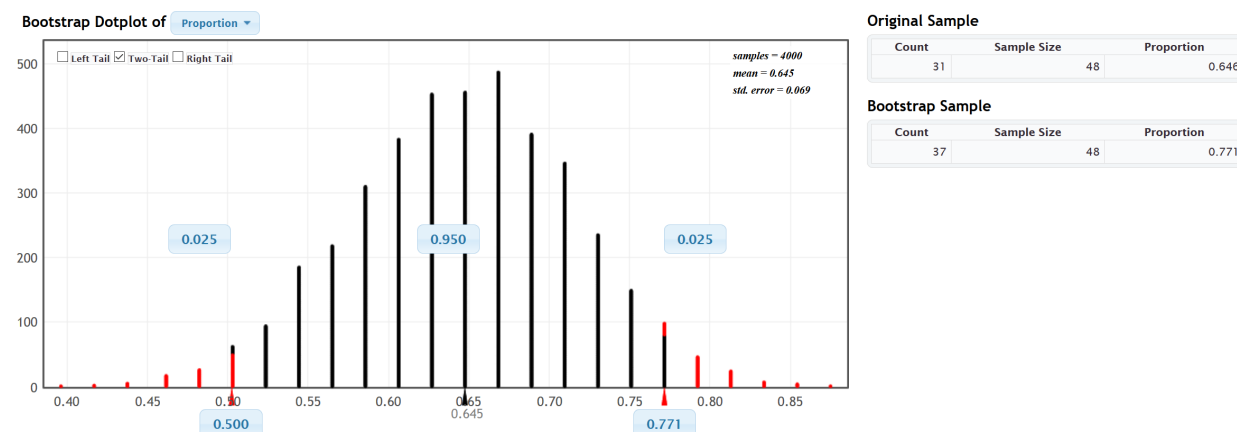
b)

The sample mean amount of carbs in these cereals of 15.043 grams could be as much as 2.061 grams off from the population mean amount of carbs in all cereals.

c)

We are 99% confident that the population mean average amount of carbs in cereals is between 12.9824 grams and 17.1036 grams.

21.



a)

Yes. Meets the assumptions for a bootstrap.

Random Sample? Yes. Given in the problem.

Individuals independent? Probably. A small sample out of a large population would probably be independent as long as the data did not rely on one machine.



b)

Answers will vary. In bootstrap above, there were 4000 bootstrap samples.

c)

Answers will vary. The bootstrap distribution above looks normal (bell shaped).

d)

Answers will vary. The confidence interval in the bootstrap distribution above is (0.500 , 0.771). They are very close to the traditional formula approach in #7.

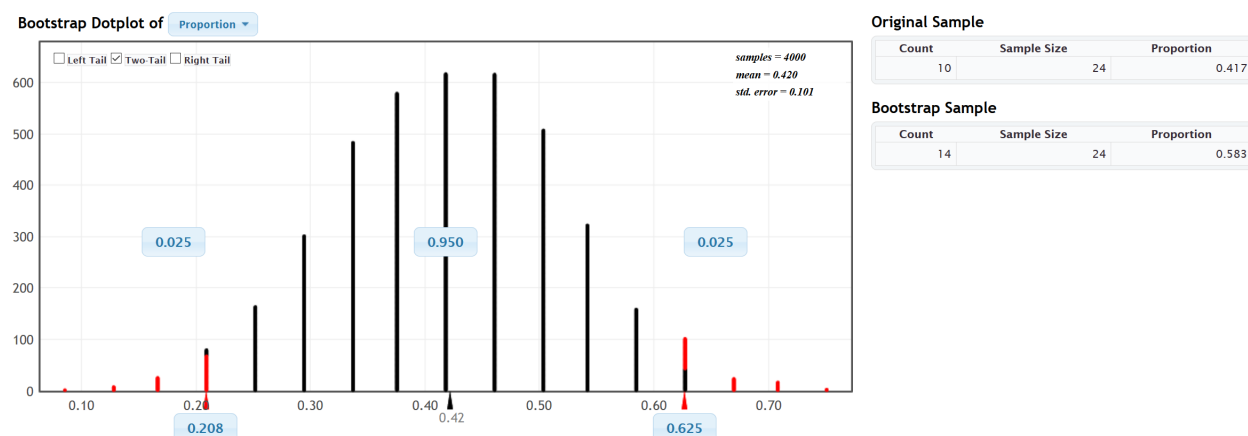
e) Answers will vary. Here are the sentences for the example above.

We are 95% confident that the population proportion of lies that are detected is between 0.500 and 0.771.

OR

We are 95% confident that the population percentage of lies that are detected is between 50.0% and 77.1%.

23.



a)

Might not meet the assumptions for a bootstrap.

Random Sample? Yes. Given in the problem.

Individuals independent? Probably not. The population is rather small and some companies be incline to put similar amounts of sugar in their cereals.

b)

Answers will vary. In bootstrap above, there were 4000 bootstrap samples.

c)

Answers will vary. The bootstrap distribution above looks normal (bell shaped).

d)

Answers will vary. The confidence interval in the bootstrap distribution above is (0.208 , 0.625). They are very close to the traditional formula approach in #11.



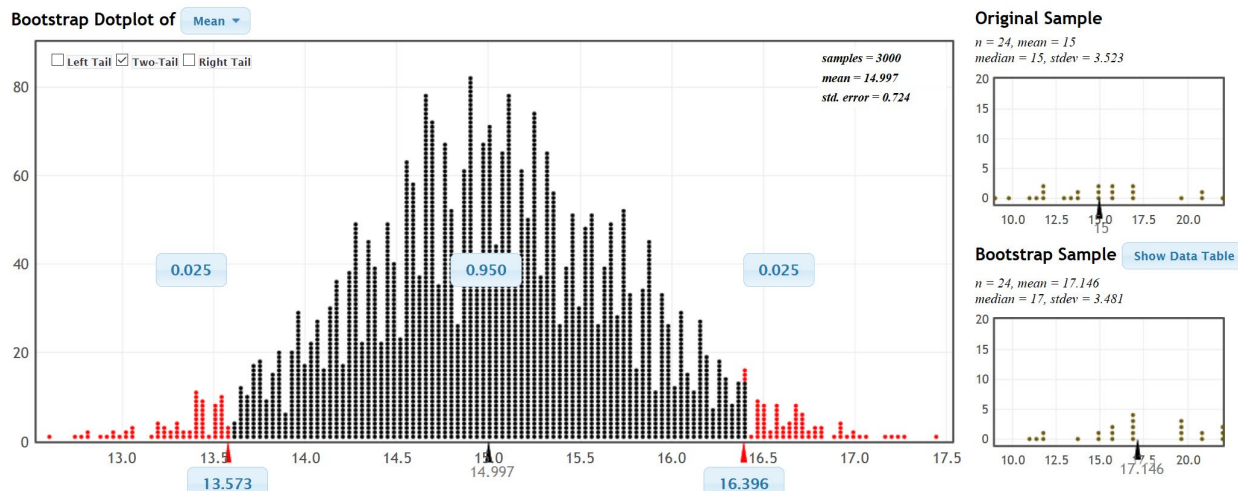
e) Answers will vary. Here are the sentences for the example above.

We are 95% confident that the population proportion of cereals with high sugar content is between 0.208 and 0.625.

OR

We are 95% confident that the population percentage of cereals with high sugar content is between 20.8% and 62.5%.

25.



a)

Might not meet the assumptions for a bootstrap.

Random Sample? Yes. Given in the problem.

Individuals independent? Probably not. The population is rather small and some companies be incline to have similar amounts of carbs in their cereals.

b)

Answers will vary. In bootstrap above, there were 3000 bootstrap samples.

c)

Answers will vary. The bootstrap distribution above looks normal (bell shaped).

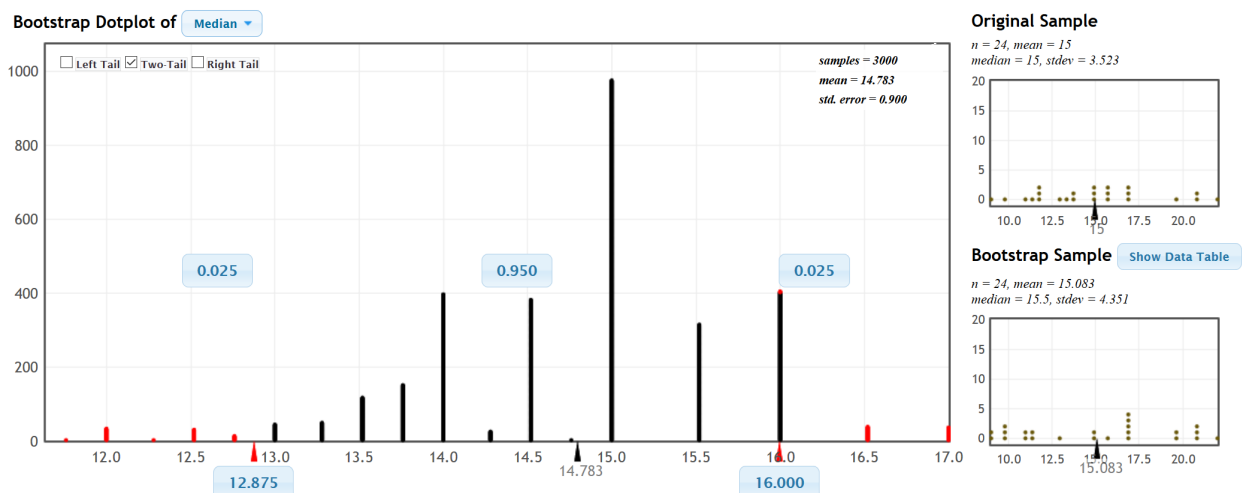
d)

Answers will vary. The confidence interval in the bootstrap distribution above is (13.573 carbs , 16.396 carbs). They are very close to the traditional formula approach in #19.

e) Answers will vary. Here are the sentences for the example above.

We are 95% confident that the population mean amount of carbs in cereals is between 13.573 carbs and 16.396 carbs.





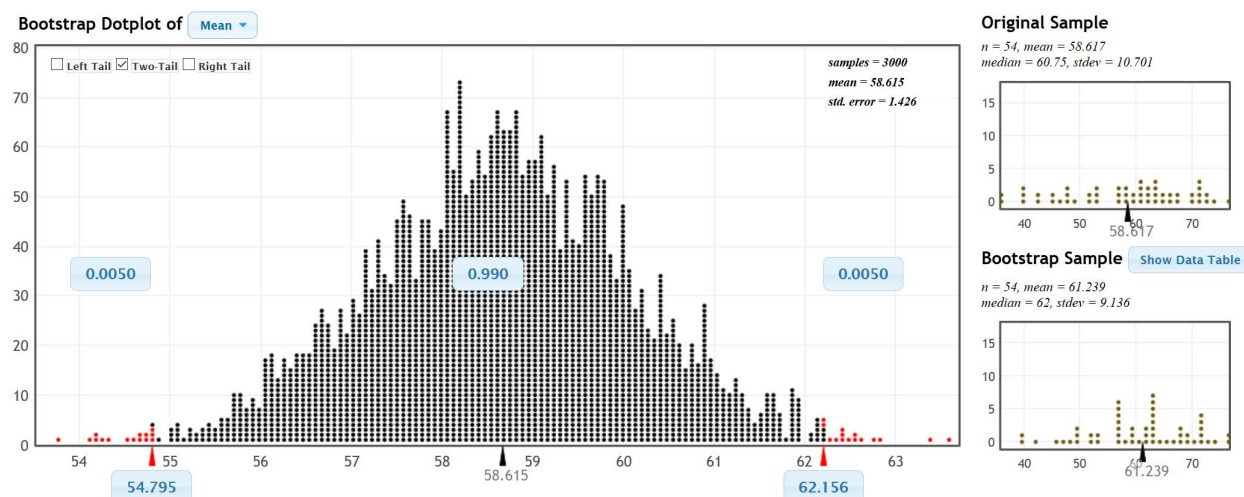
f) Answers will vary. The median bootstrap example above looks skewed left.

g) Answers will vary. The median bootstrap example above has a confidence interval of (12.875 carbs, 16.000 carbs).

h) Answers will vary. Here is the sentence for the median bootstrap example above.

We are 95% confident that the population median average amount of carbs in cereals is between 12.875 carbs and 16.000 carbs.

27.



a)

Yes. It does meet the assumptions for a bootstrap.

Random Sample? Yes. Given in the problem.

Individuals independent? Probably. The population of all bears is rather large. A random sample of 54 bears from all over will probably not be related. It would fail independence if this sample was taken from the same area.

b)

Answers will vary. In bootstrap above, there were 3000 bootstrap samples.



c)

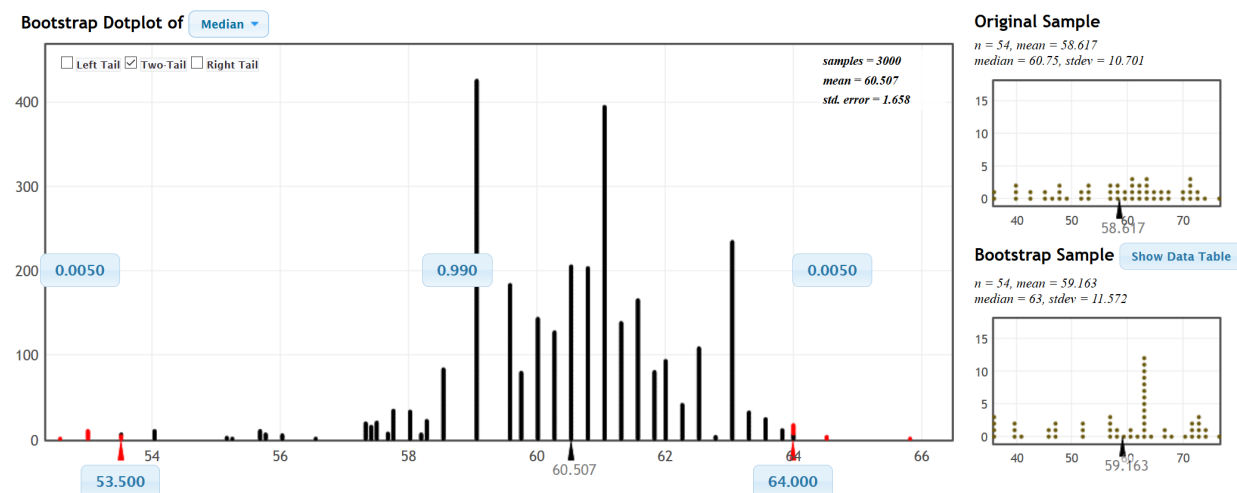
Answers will vary. The bootstrap distribution above looks normal (bell shaped).

d)

Answers will vary. The confidence interval in the bootstrap distribution above is (54.795 inches, 62.156 inches).

e) Answers will vary. Here are the sentences for the example above.

We are 99% confident that the population mean average length of this species of bear is between 54.795 inches and 62.156 inches.



f) Answers will vary. The median bootstrap example looks tri-modal.

g) Answers will vary. The median bootstrap example above has a confidence interval of (53.5 inches, 64.0 inches).

h) Answers will vary. Here is the sentence for the median bootstrap example above.

We are 95% confident that the population median average length of this species of bear is between 53.5 inches and 64.0 inches.

Section 2F Odd Answers

1.

Two-population Proportion Assumptions

- The two categorical samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- Data values between the samples should be independent of each other.
- There should be at least ten successes and at least ten failures.

2.

Two-population Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape.



Two-population Mean Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

3.

Two-population Proportion Bootstrap Assumptions

- The two categorical samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- Data values between the samples should be independent of each other.

Two-population Bootstrap Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.

Two-population Mean Bootstrap Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- Data values between the two samples should be independent of each other.

5.

a)

Population 1 is significantly lower than population 2 since the confidence interval is (negative, negative). This shows there is a negative difference between the populations.

b)

Population 1 could be between 0.068 (6.8%) and 0.115 (11.5%) lower than population 2.

c)

We are 95% confident that the population proportion for population 1 is between 0.068 and 0.115 lower than population 2.

OR

We are 95% confident that the population percentage for population 1 could be between 6.8% and 11.5% lower than population 2.

7.

a)

There is no significant difference between population 1 and population 2 since the confidence interval is (negative, positive). This means we do not know if the difference is positive or negative. Either population might be larger.

b)

We are 90% confident that there is no significant difference between the population proportions for population 1 and population 2.

OR

We are 90% confident that there is no significant difference between the population percentages for population 1 and population 2.



9.

a)

Population 1 is significantly higher than population 2 since the confidence interval is (positive, positive). This shows there is a positive difference between the populations.

b)

The population proportion for population 1 could be between 0.049 (4.9%) and 0.058 (5.8%) higher than population 2.

c)

We are 99% confident that the population proportion for population 1 is between 0.049 and 0.058 higher than population 2.

OR

We are 99% confident that the population percentage for population 1 is between 4.9% and 5.8% higher than population 2.

11.

a)

There is no significant difference between population 1 and population 2 since the confidence interval is (negative, positive). This means we do not know if the difference is positive or negative. Either population might be larger.

b)

We are 95% confident that there is no significant difference between the population proportions for population 1 and population 2.

OR

We are 95% confident that there is no significant difference between the population percentages for population 1 and population 2.

13.

a)

This data was matched pair since it was the same people measured twice.

Two-population Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population. *Yes. The data was collected randomly.*
- Data values within the sample should be independent of each other. *Yes. Since the population size is very large, a small sample of individuals from that population will probably not be related.*
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape. *Yes. The sample size was only 28, but the differences were normal. This does pass the "30 or normal" requirement.*

b)

The sample mean difference was 5.8 ACT points. Since the confidence interval was (positive, positive), this sample mean difference is significant.

c)

Since the confidence interval was (positive, positive), the ACT scores after the prep class (population 1) is significantly higher than the ACT scores before the prep class. We think that the population mean average ACT



scores after the prep class (population 1) could be between 4.4159 points and 7.1841 points higher than the ACT scores before the prep class (population 2).

d)

We are 90% confident that the population mean average ACT scores after the prep class (population 1) is between 4.4159 points and 7.1841 points higher than the ACT scores before the prep class (population 2).

15.

a)

This data was not matched pair since it was two independent groups of people.

Two-population Mean Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population. **Yes.** *The two samples were collected randomly.*
- Data values within each sample should be independent of each other. **Yes.** *Each sample was random from a large population so individuals are not likely to be related.*
- Data values between the two samples should be independent of each other. **Yes.** *Since the population is large, the people that live with smokers are not likely to be related to the people that do not live with smokers.*
- The sample sizes should be at least 30 or have a nearly normal shape. **Yes.** *Though the shapes are unknown, both sample sizes were over 30. So both samples pass the "30 or normal" requirement.*

b)

The difference between the sample means was 5.6 mg/mL. Since the confidence interval was (negative, negative), this difference is significant.

c)

Since the confidence interval was (negative, negative), the population mean average cotinine for those that do not live with smokers (population 1) is significantly lower than for those that live with smokers. We think that the population mean average cotinine level for those that do not live with smokers (population 1) could be between 18.5696 ng/mL and 24.0304 ng/mL lower than for those that live with smokers (population 2).

d)

We are 95% confident that the population mean average cotinine level for those that do not live with smokers (population 1) is between 18.5696 ng/mL and 24.0304 ng/mL lower than for those that live with smokers (population 2).

17.

a)

This data was matched pair since it was the same people measured twice.

Two-population Mean Bootstrap Assumptions (Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population. **Yes.** *The two samples were collected randomly.*
- Data values within each sample should be independent of each other. **Yes.** *Individual data was collected randomly from a large population so individuals are not likely to be related.*

b)

The sample mean difference was -43.375 mm of Hg. Since the confidence interval was (negative, negative), this difference is significant.



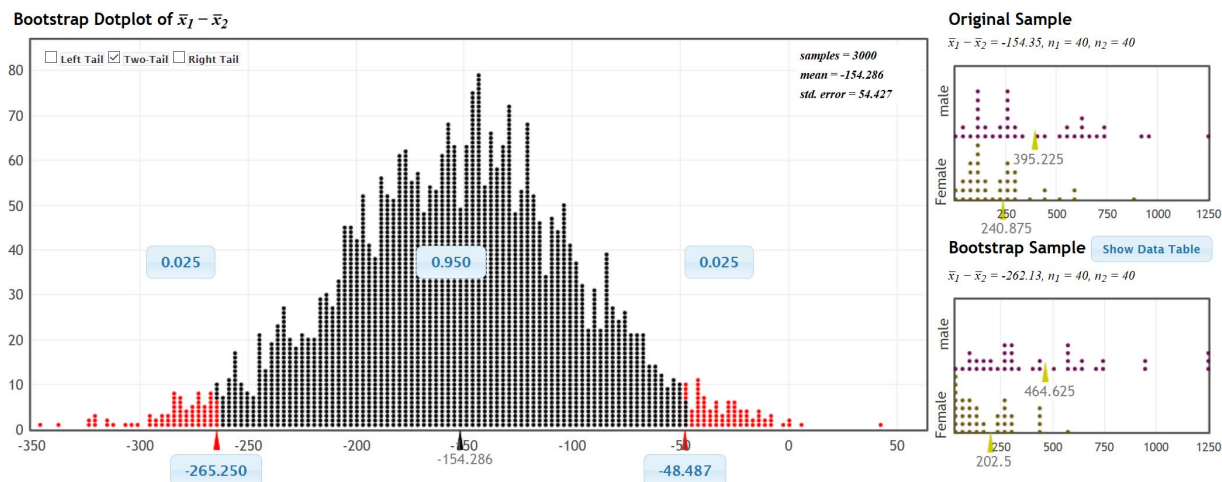
c)

Since the confidence interval was (negative, negative), the population mean average diastolic blood pressure (population 1) is significantly lower than the systolic blood pressure. We think that the population mean average diastolic blood pressure (population 1) is between 40.162 mm of Hg and 47.0 mm of Hg lower than the systolic blood pressure (population 2).

d)

We are 95% confident that the population mean average diastolic blood pressure (population 1) is between 40.162 mm of Hg and 47.0 mm of Hg lower than the systolic blood pressure (population 2).

19.



a)

This data was not matched pair since it was two independent groups of people.

Two-population Mean Bootstrap Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population. **Yes.** *The two samples were collected randomly.*
- Data values within each sample should be independent of each other. **Yes.** *Each sample was random from a large population so individuals are not likely to be related.*
- Data values between the two samples should be independent of each other. **Yes.** *Since the population is large, the females and males are not likely to be related.*

b)

Answers will vary. For the bootstrap example above, the difference between the sample means was 154.35 mg/dL. Since the confidence interval was (negative, negative), this difference is significant.

c)

Answers will vary. The confidence interval in the example above was (negative, negative), the population mean average cholesterol for women (population 1) is significantly lower than for men (population 2). We think that the population mean average cholesterol for women (population 1) could be between 48.487 mg/dL and 265.25 mg/dL lower than for men (population 2).

d)

We are 95% confident that the population mean average cholesterol for women (population 1) is between 48.487 mg/dL and 265.25 mg/dL lower than for men (population 2).



Section 2G Odd Answers

1.

One-Population Variance or Standard Deviation Assumptions

- The quantitative sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- The sample data must be normal.

3.

a)

The sample data does not pass the assumptions for a variance or standard deviation confidence interval.

One-Population Variance or Standard Deviation Assumptions

- The quantitative sample data should be collected randomly or be representative of the population. **Yes.** The data was collected randomly. This was given in #2.
- Data values within the sample should be independent of each other. **Yes.** Since this is a small random sample from a large population, individuals are unlikely to be related.
- The sample data must be normal. **No.** The data was skewed left.

b)

We are 90% confident that the population variance for ACT scores is between 70.8423 and 143.8391. These numbers may not be very accurate since the data did not meet the assumptions.

c)

We are 90% confident that the population standard deviation for ACT scores is between 8.4168 ACT points and 11.9933 ACT points. These numbers may not be very accurate since the data did not meet the assumptions.

5.

a)

The sample data does pass the assumptions for both variance and standard deviation confidence intervals.

One-Population Variance or Standard Deviation Assumptions

- The quantitative sample data should be collected randomly or be representative of the population. **Yes.** The data was collected randomly. This was given in #4.
- Data values within the sample should be independent of each other. **Yes.** Since this is a small random sample from a large population, individuals are unlikely to be related.
- The sample data must be normal. **Yes.** The sample data was normal (bell shaped).

b)

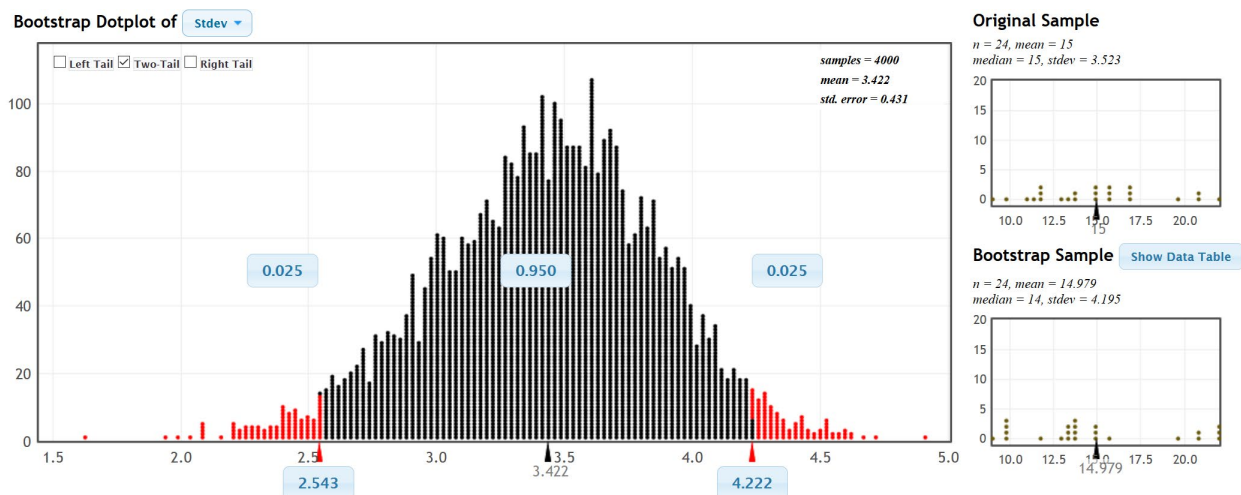
We are 99% confident that the population variance for human body temperature is between 0.3666 and 1.0524.

c)

We are 99% confident that the population standard deviation for human body temperature is between 0.6054 °F and 1.0258 °F.



7.



a)

The sample data may not pass the assumptions for standard deviation bootstrap confidence intervals.

One-Population Variance or Standard Deviation Assumptions

- The quantitative sample data should be collected randomly or be representative of the population. **Yes.** *The data was collected randomly.*
- Data values within the sample should be independent of each other. **Probably not.** *The population size is rather small and individual cereals may be related to each other since they are made by the same companies.*

b)

Answers will vary. The example bootstrap above had 4000 bootstrap samples.

c)

Answers may vary. The example bootstrap distribution above looks normal (bell shaped).

d)

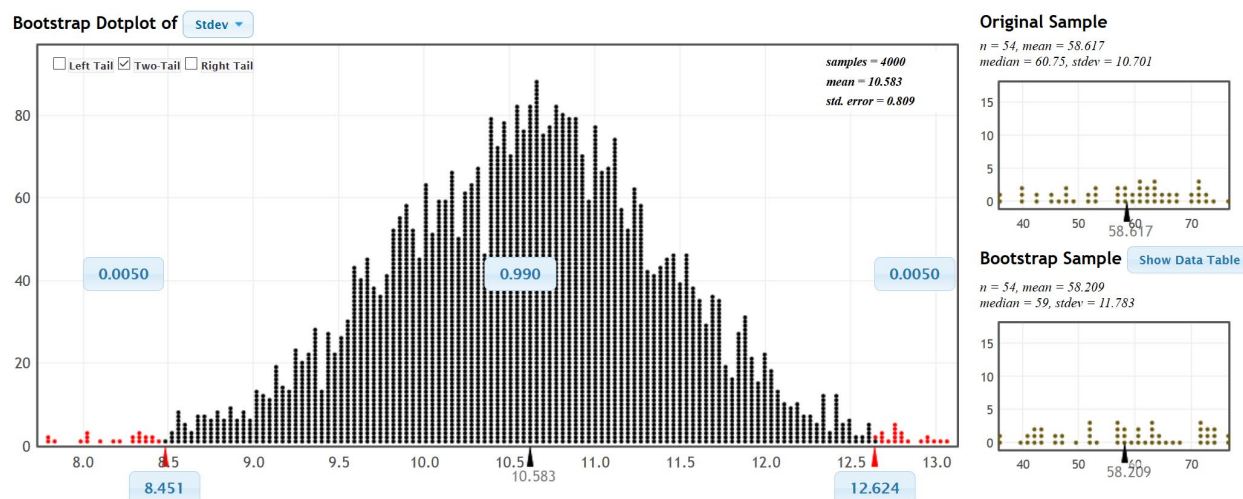
Answers will vary. The example bootstrap confidence interval above was (2.543 carbs, 4.222 carbs).

e)

We are 95% confident that the population standard deviation for the amount of carbs in cereals is between 2.543 carbs and 4.222 carbs. These numbers may not be accurate since the sample data did not pass the assumptions.



9.



a)

The sample data does pass the assumptions for standard deviation bootstrap confidence intervals.

One-Population Variance or Standard Deviation Assumptions

- The quantitative sample data should be collected randomly or be representative of the population. **Yes.** The data was collected randomly.
- Data values within the sample should be independent of each other. **Probably.** Since it was a small random sample from a large population from different areas the bears are likely to be independent of each other.

b)

Answers will vary. The example bootstrap above had 4000 bootstrap samples.

c)

Answers may vary. The example bootstrap distribution above looks normal (bell shaped).

d)

Answers will vary. The example bootstrap confidence interval above was (8.451 inches, 12.624 inches).

e)

Answers will vary. Here is the sentence for the example bootstrap above.

We are 99% confident that the population standard deviation for the lengths of this species of bear is between 8.451 inches and 12.624 inches.

Chapter 2 Review All Answers

1.

N: Parameter describing the size of a population.

n: Statistic describing the size of a sample.

π or p : Parameter describing a population proportion.



\hat{p} : Statistic describing a sample proportion.

μ : Parameter describing a population mean average.

\bar{x} : Statistic describing a sample mean average.

σ : Parameter describing a population standard deviation.

s : Statistic describing a sample standard deviation.

σ^2 : Parameter describing a population variance.

s^2 : Statistic describing a sample variance.

ρ : Parameter describing a population correlation coefficient.

r : Statistic describing the correlation coefficient from sample data.

β_1 : Parameter describing a population slope.

b_1 : Statistic describing a slope from sample data.

2.

a)

$\bar{x} = 101.9$ (statistic)

$s = 14.8$ (statistic)

$s^2 = 219.04$ (statistic)

$\mu = 100$ (parameter)

$\sigma = 15$ (parameter)

$\sigma^2 = 225$ (parameter)

b)

$\rho = 0$ (parameter)

$\beta_1 =$

$\beta_1 = 20$ (parameter)

$r = 0.338$ (statistic)

$b_1 = 13.79$ (statistic)

c)

$\hat{p} = 0.03$ (statistic)

$\pi = 0.015$ (parameter)

d)

$n = 238$ (statistic)

$N = 5,000,000$ (parameter)

3.

One-population Mean Assumptions

- The quantitative sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- The sample size should be at least 30 or have a nearly normal shape.



4.

One-Population Variance or Standard Deviation Assumptions

- The quantitative sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- The sample data must be normal.

5.

One-population Proportion Assumptions

- The categorical sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least ten successes and at least ten failures.

6.

Two-population Mean Assumptions (Matched Pair)

- The quantitative ordered pair sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape.

Two-population Mean Assumptions (Not Matched Pair)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

7.

Two-population Proportion Assumptions

- The two categorical samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- Data values between the samples should be independent of each other.
- There should be at least ten successes and at least ten failures.

8.

Bootstrap and Randomized Simulation Assumptions

- The sample data should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- If multiple samples were collected that were not matched pair, then the data values between the samples should be independent of each other.

9.

Vocabulary

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data.
For example, a sample mean average, a sample standard deviation, or a sample percentage.



Parameter: A number that describes the characteristics of a population like a population mean or a population percentage. Can be calculated from an unbiased census, but is often just a guess about the population.

Sampling Distribution: Take many random samples from a population, calculate a sample statistic like a mean or percent from each sample and graph all of the sample statistics on the same graph. The center of the sampling distribution is a good estimate of the population parameter.

Sampling Variability: Random samples values and sample statistics are usually different from each other and usually different from the population parameter.

Point Estimate: When someone takes a sample statistic and then claims that it is the population parameter.

Margin of Error: Total distance that a sample statistic might be from the population parameter. For normal sampling distributions and a 95% confidence interval, the margin of error is approximately twice as large as the standard error.

Standard Error: The standard deviation of a sampling distribution. The distance that typical sample statistics are from the center of the sampling distribution. Since the center of the sampling distributions is usually close to the population parameter, the standard error tells us how far typical sample statistics are from the population parameter.

Confidence Interval: Two numbers that we think a population parameter is in between. Can be calculated by either a bootstrap distribution or by adding and subtracting the sample statistic and the margin of error.

95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.

90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.

99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.

Bootstrapping: Taking many random samples values from one original real random sample with replacement.

Bootstrap Sample: A simulated sample created by taking many random samples values from one original real random sample with replacement.

Bootstrap Statistic: A statistic calculated from a bootstrap sample.

Bootstrap Distribution: Putting many bootstrap statistics on the same graph in order to simulate the sampling variability in a population, calculate standard error, and create a confidence interval. The center of the bootstrap distribution is the original real sample statistic.

10.

a)

We are 99% confident that the population mean weight is between 55.6 pounds and 69.4 pounds.

b)

We are 90% confident that the population proportion is in between 0.352 and 0.411.

OR

We are 90% confident that the population percentage is in between 35.2% and 41.1%.

c)

We are 95% confident that the population standard deviation is between 3.1 pounds and 4.7 pounds.



d)

We are 99% confident that the population variance is between 461.8 square inches and 591.3 square inches.

e)

Since the confidence interval is (positive, positive), this indicates that population 1 is significantly larger than population 2. The confidence interval indicates that population 1 is between 13.2 kg and 14.8 kg larger.

Sentence: We are 95% confident that the population mean average weight of population 1 is between 13.2 kg and 14.8 kg larger than population 2.

f)

Since the confidence interval is (negative, positive), this indicates that there is no significant difference between the population mean averages for population 1 and population 2. They are very close and we cannot tell which population is larger.

Sentence: We are 90% confident that there is no significant difference between the population mean averages for population 1 and population 2.

g)

Since the confidence interval is (negative, positive), this indicates that there is no significant difference between the population proportions for population 1 and population 2. The population percentages are very close and we cannot tell which population is larger.

Sentence: We are 95% confident that there is no significant difference between the population proportions for population 1 and population 2.

OR

We are 95% confident that there is no significant difference between the population percentages for population 1 and population 2.

h)

Since the confidence interval is (negative, negative), this indicates that population 1 is significantly smaller than population 2. The confidence interval indicates that population 1 is between 0.057 (5.7%) and 0.072 (7.2%) smaller than population 2.

Sentence: We are 99% confident that the population proportion for population 1 is between 0.057 and 0.072 less than population 2.

OR

We are 99% confident that the population percentage for population 1 is between 5.7% and 7.2% less than population 2.

11.

A sampling distribution is created by taking hundreds or thousands of random samples from a population. We then calculate a sample statistic like the mean, standard deviation or proportion and put all of the thousands of random sample statistics on the same graph. The center of the sampling distribution is often very close to the actual population parameter. The standard error is the standard deviation of the sampling distribution and helps us understand how far typical statistics may be from the center (population parameter). Overall the sampling distribution teaches us about sampling variability. This is the principle that random samples are usually very different from each other and are often very far off from the population parameter.

12.

If the sampling distribution is normal, then we can multiply the critical value Z-score or T-score times the standard error to get the margin of error. If we take the original random sample statistic and add and subtract the margin of error, we will get the upper and lower limits of the confidence interval.



13.

$$\frac{s^2(n-1)}{\chi^2_{upper}} < \text{Population Variance } (\sigma^2) < \frac{s^2(n-1)}{\chi^2_{lower}}$$

To create a confidence interval for the population variance, multiply the sample variance s^2 by the degrees of freedom $n-1$ and then divide by the chi-squared critical values.

14.

a)

William Gosset

b)

Guinness Beer

c)

Needed better accuracy for smaller data sets.

d)

Guinness Beer had a strict no publishing policy and would have fired him.

e)

"student"

f)

When the sample size is small, the T-scores are significantly greater than the Z-scores. This accounts for more variability in smaller data sets.

g)

When the sample size is large, the T-scores and Z-scores are about the same.

h)

One and two-population proportion (percentage) confidence intervals use Z-scores.

i)

One and two-population mean average confidence intervals use T-scores.

j)

For one quantitative data set with a sample size "n", the degrees of freedom is "n - 1".

15.

The central limit theorem states that a sampling distribution made of sample means will be normal (bell shaped) if the sample size is sufficiently large. The mean and standard deviation of a sampling distribution are important calculations in inferential statistics but are only accurate if the sampling distribution is normal. The central limit theorem discusses what conditions need to be met for the sampling distributions to be normal. These are the foundations of our assumptions for the confidence intervals. For quantitative sample data that is not normal, we will need a sample size of at least 30 to ensure the sampling distribution will look nearly normal. For quantitative sample data that is already normal, we can use a sample below 30. For categorical data, we will need at least 10 successes and failures for the sampling distribution for proportions to look normal.



16.

Suppose we want to make a confidence interval, but the data does not pass the assumptions for our sampling distribution to look normal. This would mean our traditional formulas involving standard error, Z-scores and T-scores may not be very accurate. Bootstrapping is a technique for calculating a confidence interval directly without the traditional formula. In a bootstrap, we take thousands of random samples with replacement from the one random sample. We then calculate the statistic from each of the bootstrap samples and put all of the thousand bootstrap statistics on the bootstrap distribution. In a sense, we have created a simulated sampling distribution but not from the population. Find the middle 95%, 90% or 99% markers from the distribution and you have your confidence interval. Bootstrapping still requires the data to be collected randomly and individual observations should be independent, but it does not require the same central limit theorem assumptions. The bootstrap distribution does not have to look normal since we are calculating the middle 95%, 90% or 99% directly.

Section 3A Odd Answers

1.

$$H_0: \pi = 0.93$$

$$H_A: \pi < 0.93 \text{ (claim)}$$

Left-tailed test

3.

$$H_0: \pi = 0.74$$

$$H_A: \pi > 0.74 \text{ (claim)}$$

Right-tailed test

5.

$$H_0: \sigma = 2.9 \text{ inches}$$

$$H_A: \sigma \neq 2.9 \text{ inches (claim)}$$

Two-tailed test

7.

 π_1 : The population proportion of women that hold CEO level jobs.

 π_2 : The population proportion of men that hold CEO level jobs.

$$H_0: \pi_1 = \pi_2$$

$$H_A: \pi_1 < \pi_2 \text{ (claim)}$$

Left-tailed test

9.

 π : The percentage of republicans that support decreasing taxes.

$$H_0: \pi = 0.5$$

$$H_A: \pi > 0.5 \text{ (claim)}$$

Right-tailed test



11.

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 > 0 \text{ (claim)}$$

Right-tailed test

13.

$$H_0: \mu = 18 \text{ thousand dollars}$$

$$H_A: \mu < 18 \text{ thousand dollars (claim)}$$

Left-tailed test

15.

$$H_0: \sigma^2 = 0.5$$

$$H_A: \sigma^2 > 0.5 \text{ (claim)}$$

Right-tailed test

17.

 μ_1 : The population mean average salary of female lawyers in NY.

 μ_2 : The population mean average salary of male lawyers in NY.

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 < \mu_2 \text{ (claim)}$$

Left-tailed test

19.

$$H_0: \rho = 0$$

$$H_A: \rho > 0 \text{ (claim)}$$

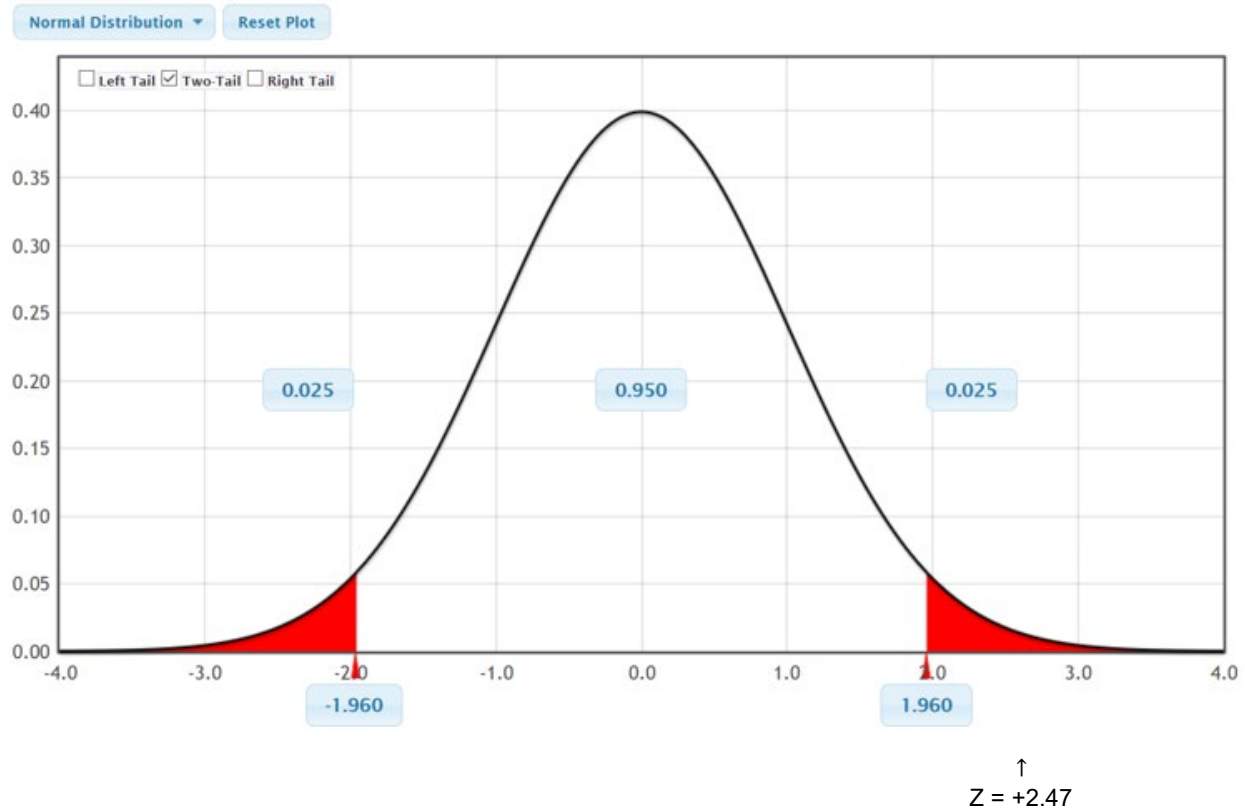
Right-tailed test



Section 3B Odd Answers

1.

a)



b)

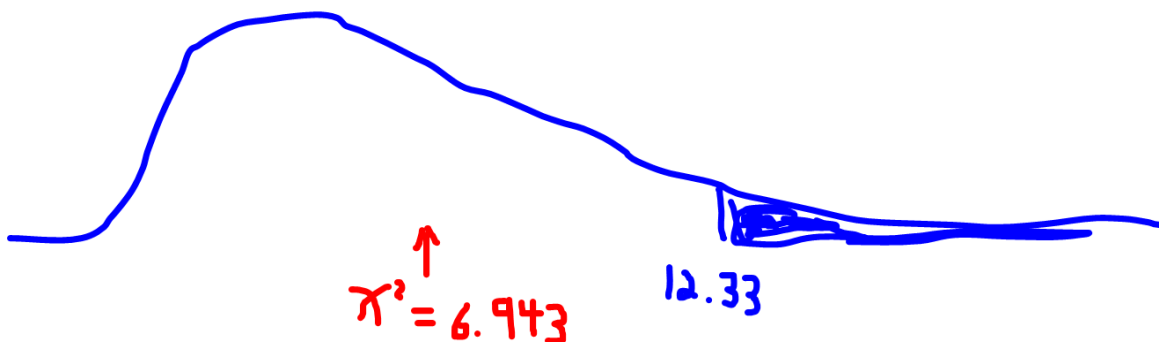
The test statistic does fall in the right tail of the normal Z-score distribution.

c)

Since the test statistic fell in the tail determined by the critical values, the sample data does significantly disagree with the null hypothesis.

3.

a)



b)

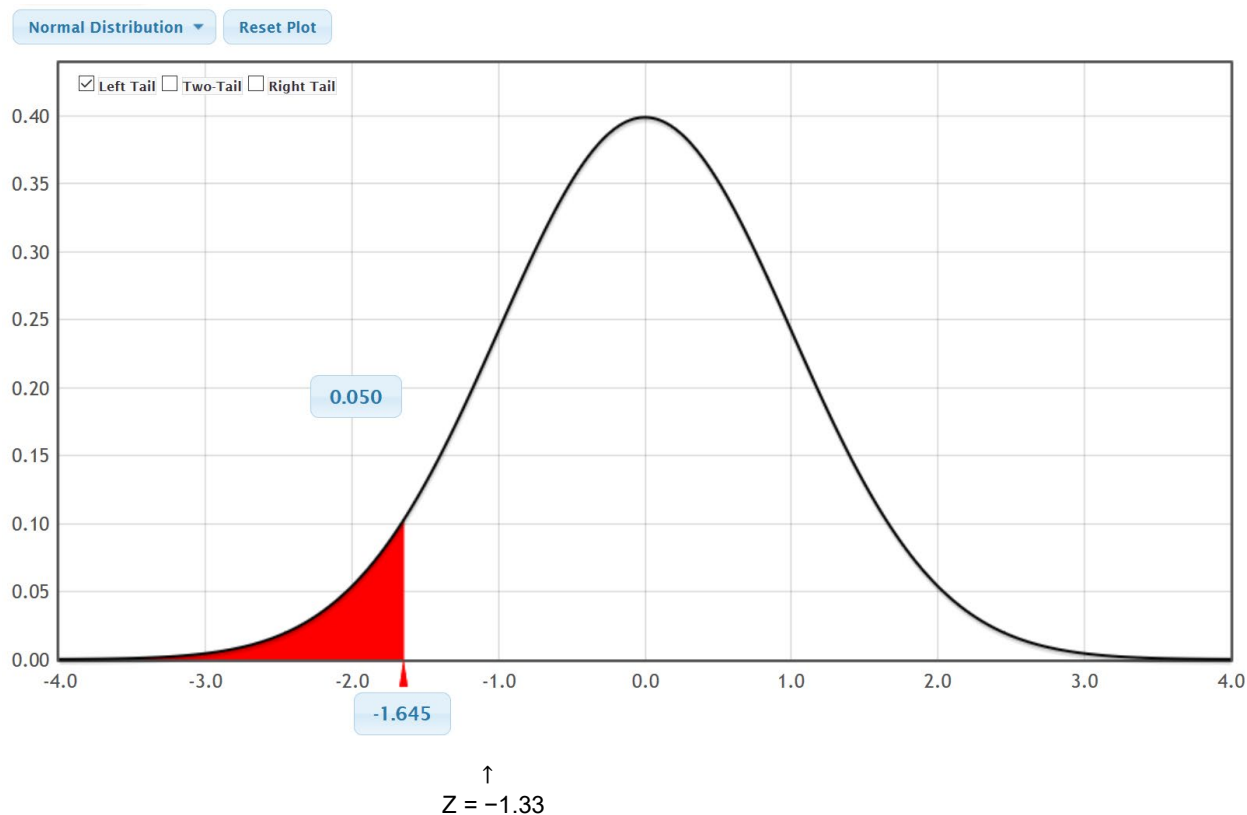
The test statistic does NOT fall in the right tail of the Chi-squared distribution.

c)

Since the test statistic did NOT fall in the tail determined by the critical values, the sample data does NOT significantly disagree with the null hypothesis.

5.

a)



b)

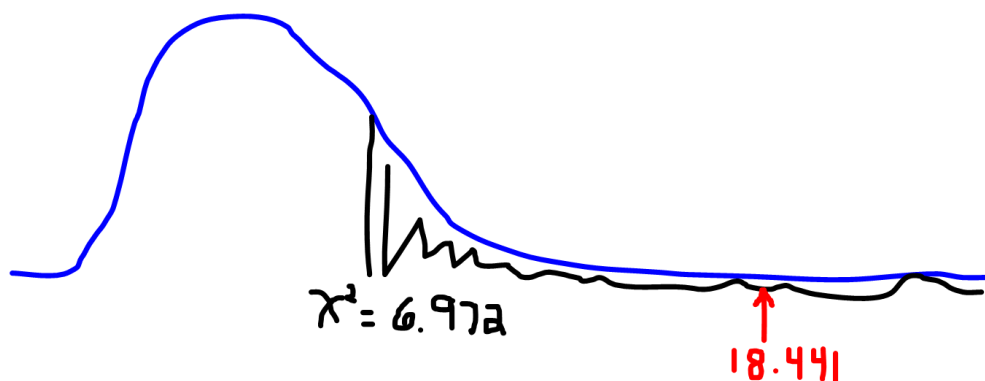
The test statistic does NOT fall in the left tail of the normal Z-score distribution.

c)

Since the test statistic did NOT fall in the tail determined by the critical values, the sample data does NOT significantly disagree with the null hypothesis.

7.

a)



b)

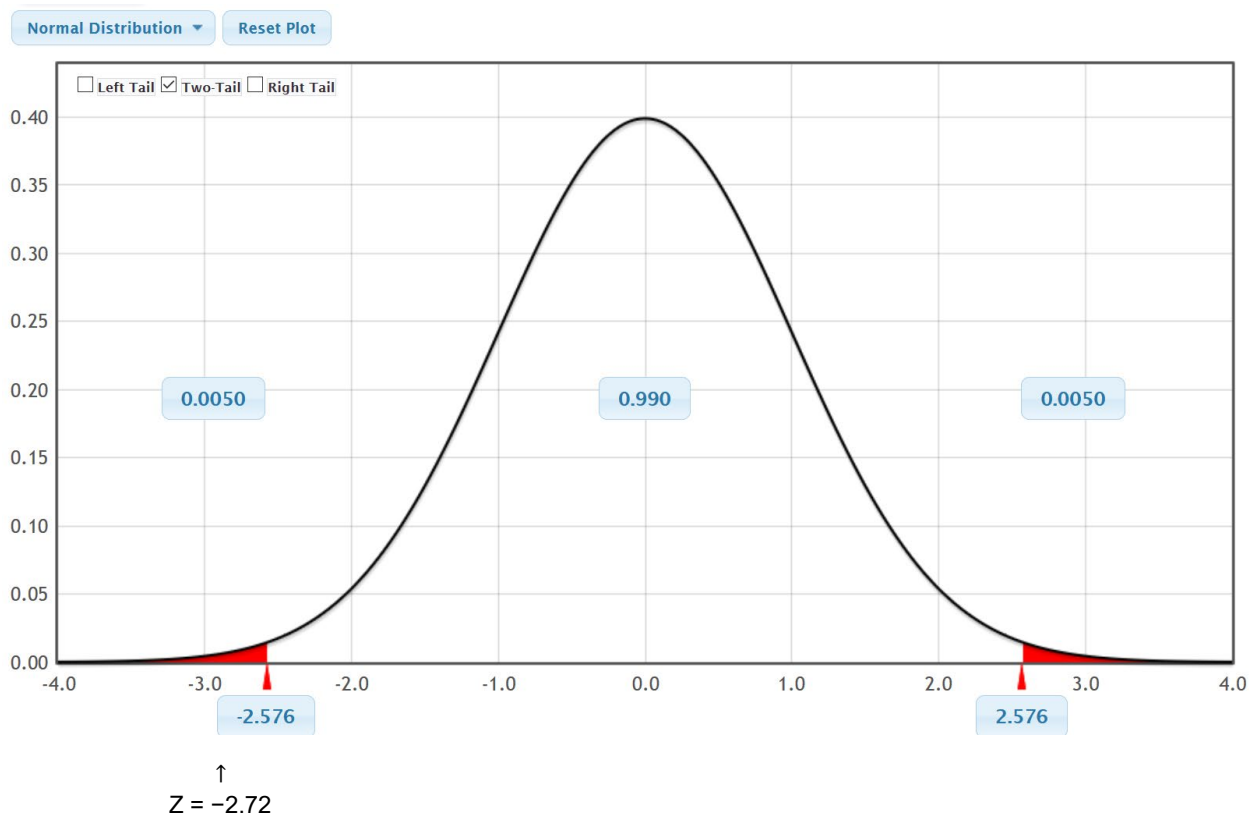
The test statistic does fall in the right tail of the Chi-squared distribution.

c)

Since the test statistic fell in the tail determined by the critical values, the sample data does significantly disagree with the null hypothesis.

9.

a)



b)

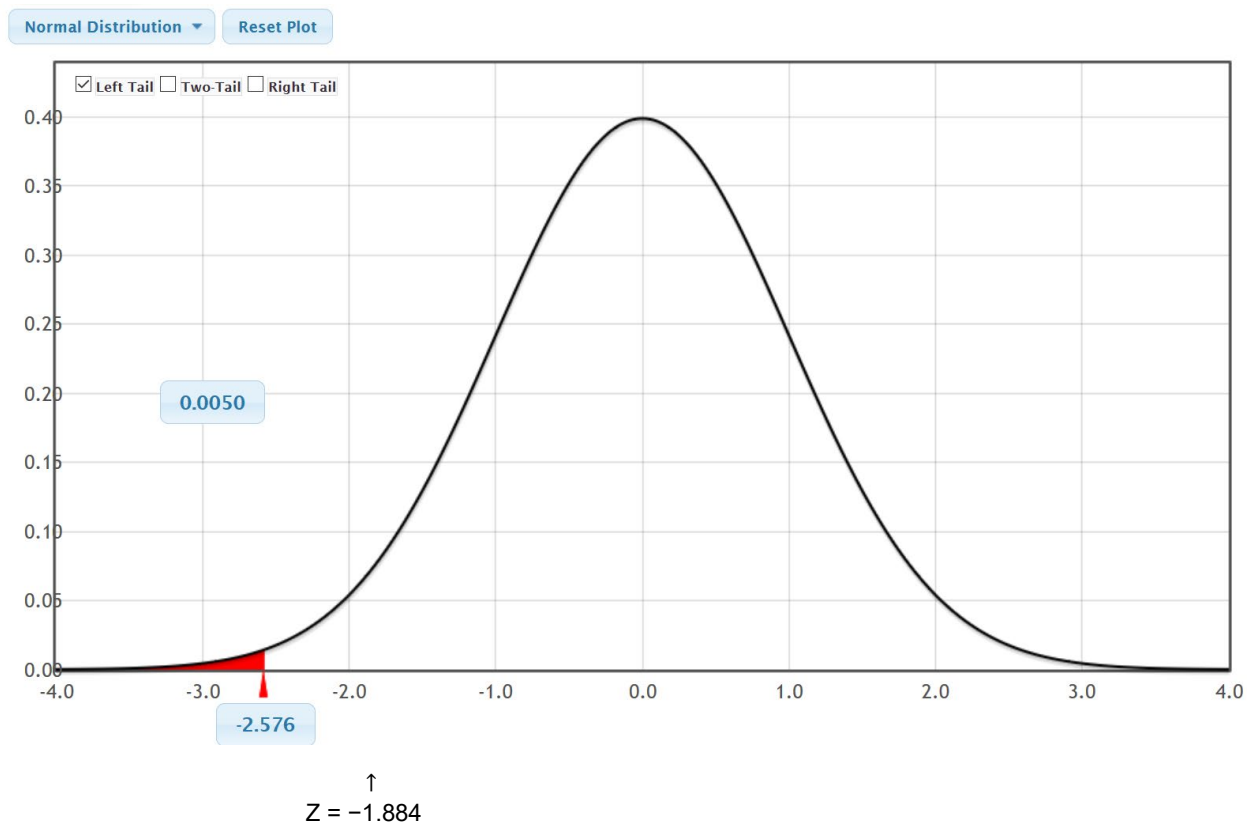
The test statistic does fall in the left tail of the normal Z-score distribution.

c)

Since the test statistic fell in one of the tails determined by the critical values, the sample data does significantly disagree with the null hypothesis.

11.

a)



b)

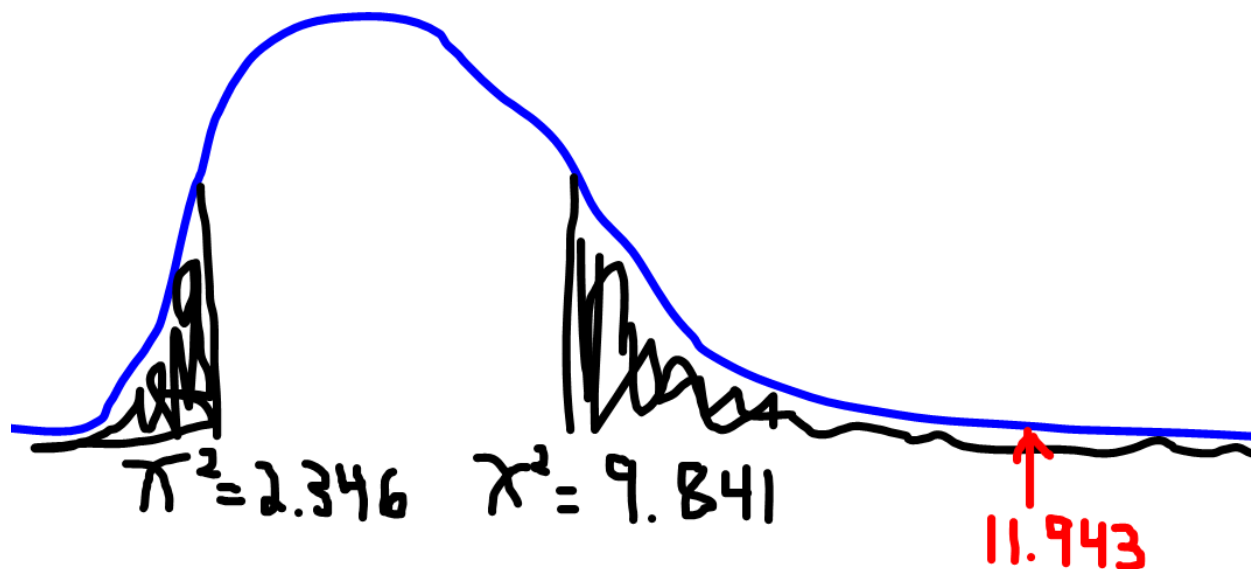
The test statistic does NOT fall in the left tail of the normal Z-score distribution.

c)

Since the test statistic did NOT fall in the left tail determined by the critical value, the sample data does NOT significantly disagree with the null hypothesis.

13.

a)



b)

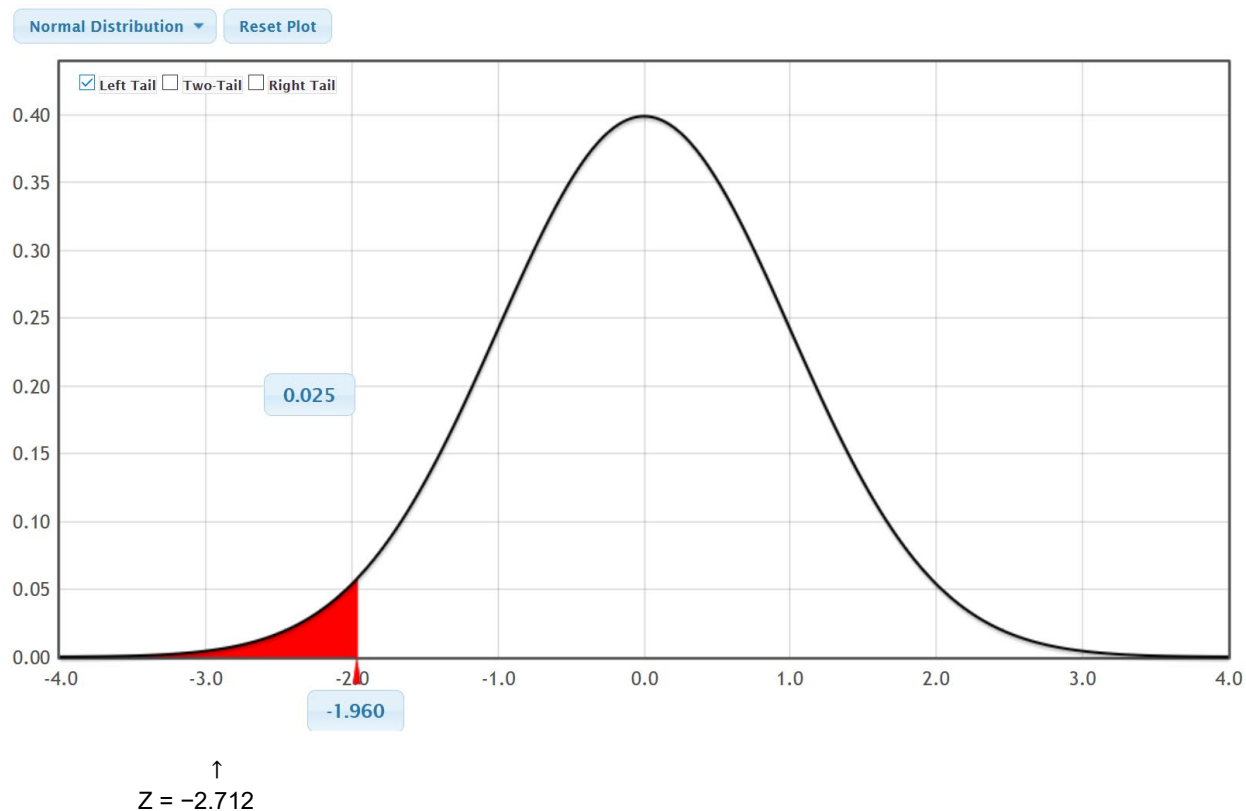
The test statistic does fall in one of the tails of the Chi-squared distribution.

c)

Since the test statistic fell in one of the tails determined by the critical values, the sample data does significantly disagree with the null hypothesis.

15.

a)



b)

The test statistic does fall in the left tail of the normal Z-score distribution.

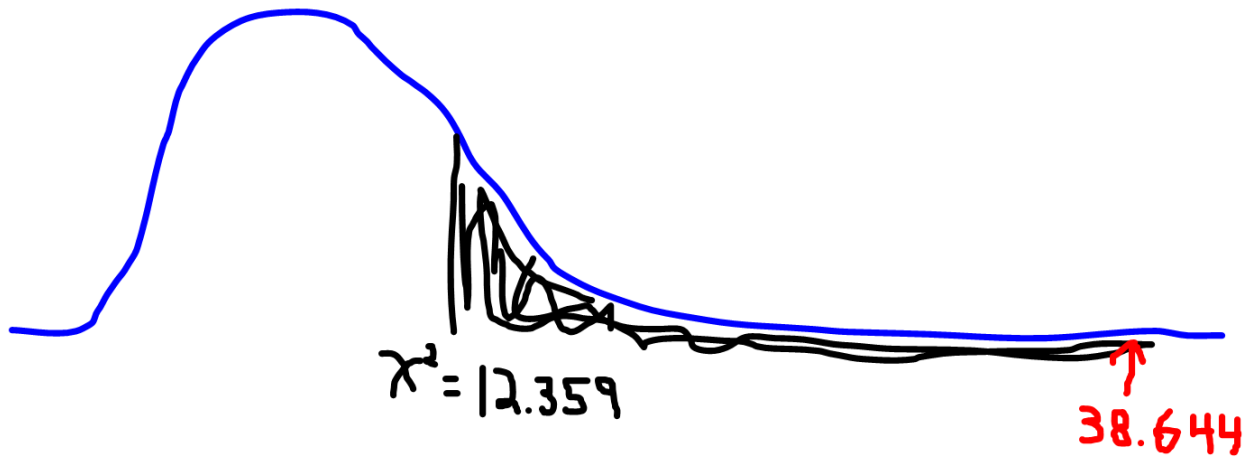
c)

Since the test statistic fell in the left tail determined by the critical value, the sample data significantly disagrees with the null hypothesis.



17.

a)



b)

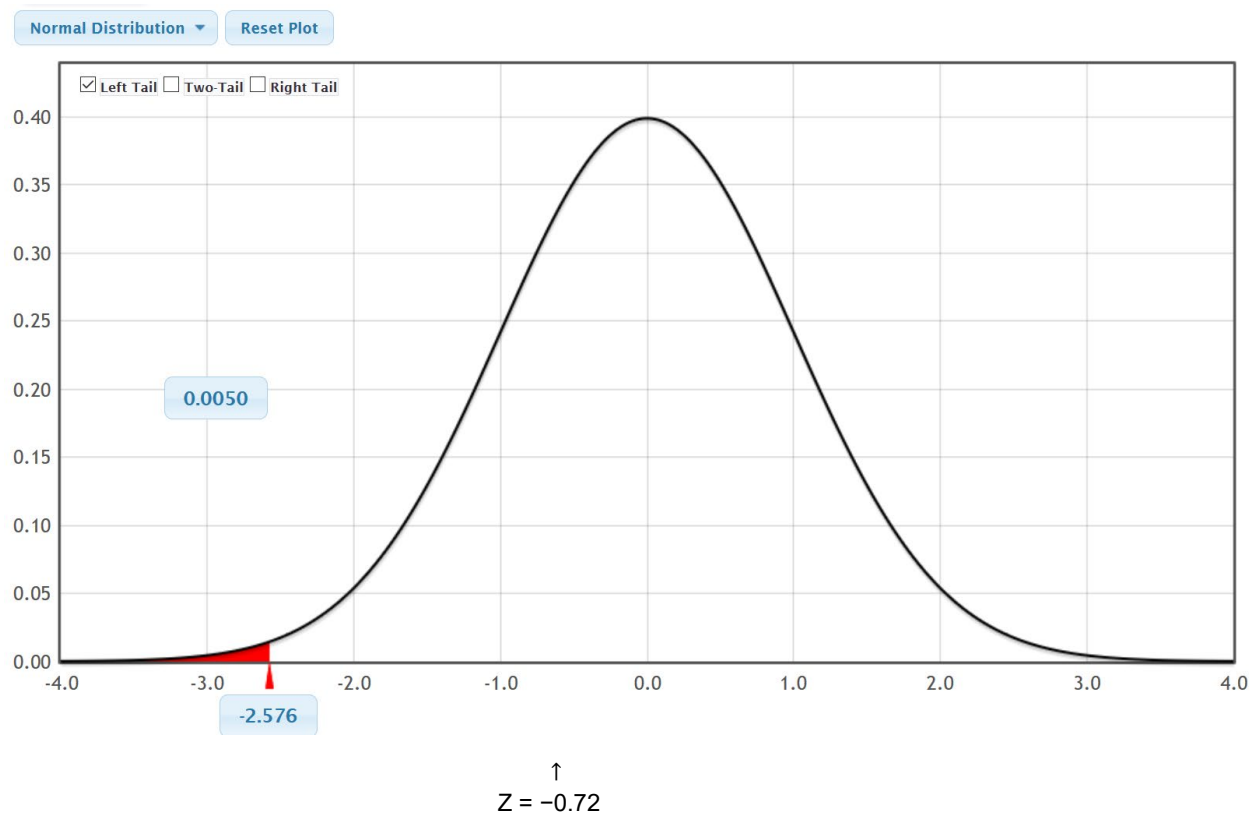
The test statistic does fall in the right tail of the Chi-squared distribution.

c)

Since the test statistic fell in the tail determined by the critical values, the sample data does significantly disagree with the null hypothesis.

19.

a)



b)

The test statistic does NOT fall in the left tail of the normal Z-score distribution.

c)

Since the test statistic did NOT fall in the left tail determined by the critical value, the sample data does NOT significantly disagree with the null hypothesis.



21.

Critical Values = ± 1.96

Since the Z-test statistic fell in one of the tails determined by the critical values, the sample data does significantly disagree with the null hypothesis.

23.

Critical Value = $+1.282$

Since the Z-test statistic fell in the right tail determined by the critical value, the sample data does significantly disagree with the null hypothesis.

25.

Critical Value = $+1.307$

Since the T-test statistic fell in the right tail determined by the critical value, the sample data does significantly disagree with the null hypothesis.

27.

Critical Value = $+42.558$

Since the chi-squared test statistic did not fall in the right tail determined by the critical value, the sample data does not significantly disagree with the null hypothesis.

29.

Critical Values: Lower critical value = $+6.843$, Upper critical value = 38.583

Since the chi-squared test statistic fell in one of the tails determined by the critical values, the sample data does significantly disagree with the null hypothesis.

31.

Z test statistic = $(0.835 - 0.9) \div 0.053 \approx -1.2264$

The sample proportion of 0.835 is 1.2264 standard errors below the population proportion of 0.9.

OR

The sample percentage of 83.5% is 1.2264 standard errors below the population percentage of 90%.

33.

T test statistic = $(135.7 - 100) \div 23.9 \approx +1.494$

The sample mean of 135.7 mg is 1.494 standard errors above the population mean of 100 mg.

35.

T test statistic = $(52.71 - 60) \div 6.42 \approx -1.136$

The sample mean of 52.71 thousand dollars is 1.136 standard errors below the population mean of 60 thousand dollars.



Section 3C Odd Answers

	<i>P</i> -value	<i>P</i> -value %	Significance Level %	Significance Level Proportion	Low <i>P</i> -value or High <i>P</i> -value?	Sample Data significantly disagree with H_0 ? (Yes or No)	Could be sampling variability or Unlikely?	Reject H_0 or fail to reject?
1.	0.238	23.8%	5%	0.05	High	Not Sig	Could be	Fail
2.	0.0003		1%					
3.	5.7×10^{-6}	0.00057%	10%	0.1	Low	Sig	Unlikely	Reject
4.	0.441		5%					
5.	0.138	13.8%	1%	0.01	High	Not Sig	Could be	Fail
6.	0		10%					
7.	0.043	4.3%	5%	0.05	Low	Sig	Unlikely	Reject
8.	0.085		1%					
9.	1.4×10^{-4}	0.014%	10%	0.1	Low	Sig	Unlikely	Reject
10.	0.112		5%					
11.	0	0%	1%	0.01	Low	Sig	Unlikely	Reject
12.	0.539		10%					
13.	0.0006	0.06%	10%	0.1	Low	Sig	Unlikely	Reject
14.	2.5×10^{-7}		1%					
15.	0.861	86.1%	5%	0.05	High	Not Sig	Could be	Fail
16.	0.199		5%					

	<i>P</i> -value	<i>P</i> -value %	Significance Level %	Significance Level Proportion	Low <i>P</i> -value or High <i>P</i> -value?	Sample Data significantly disagree with H_0 ? (Yes or No)	Could be sampling variability or Unlikely?	Reject H_0 or fail to reject?
17.	0.034	3.4%	5%	0.05	Low	Sig	Unlikely	Reject
18.	0.128		1%					
19.	8.6×10^{-4}	0.086%	10%	0.1	Low	Sig	Unlikely	Reject
20.	0.0437		5%					
21.	0	0%	1%	0.01	Low	Sig	Unlikely	Reject
22.	0.612		10%					
23.	0.087	8.7%	5%	0.05	High	Not Sig	Could be	Fail
24.	0.0048		1%					
25.	5.5×10^{-7}	0.000055%	10%	0.1	Low	Sig	Unlikely	Reject
26.	0.0216		5%					
27.	0.444	44.4%	1%	0.01	High	Not Sig	Could be	Fail
28.	0.0539		10%					
29.	0.722	72.2%	10%	0.1	High	Not Sig	Could be	Fail
30.	3.8×10^{-3}		1%					
31.	0.0823	8.23%	5%	0.05	High	Not Sig	Could be	Fail
32.	0.0227		5%					

33.

a) 0.0168 or 1.68%

b) If the null hypothesis is true, there is a 1.68% probability of getting the sample data or more extreme because of sampling variability.

c) Since the *P*-value 1.68% was lower than the significance level of 5%, the sample data does significantly disagree with the population proportion 0.93 in the null hypothesis.



d) Since the P-value 1.68% was lower than the significance level of 5%, it is unlikely for this sample data to occur because of sampling variability.

e) Since the P-value 1.68% was lower than the significance level of 5%, we should reject the null hypothesis.

35.

a) 0.1244 or 12.44%

b) If the null hypothesis is true, there is a 12.44% probability of getting the sample data or more extreme because of sampling variability.

c) Since the P-value 12.44% was higher than the significance level of 5%, the sample data does not significantly disagree with the population proportion 0.74 in the null hypothesis.

d) Since the P-value 12.44% was higher than the significance level of 5%, this sample data could have occurred because of sampling variability.

e) Since the P-value 12.44% was higher than the significance level of 5%, we should fail to reject the null hypothesis.

37.

a) 0.895 or 89.5%

b) If the null hypothesis is true, there is an 89.5% probability of getting the sample data or more extreme because of sampling variability.

c) Since the P-value 89.5% was higher than the significance level of 10%, the sample data does not significantly disagree with the population proportion 0.1 in the null hypothesis.

d) Since the P-value 89.5% was higher than the significance level of 10%, this sample data could have occurred because of sampling variability.

e) Since the P-value 89.5% was higher than the significance level of 10%, we should fail to reject the null hypothesis.

39.

P-value = 0.084 or 8.4%

41.

P-value = $0.0098 + 0.0098 = 0.0196$ or 1.96%

43.

P-value = 0.053 or 5.3%

45.

P-value = 0.000018 or 0.0018%

Section 3D Odd Answers

	Claim	P-value	Evidence (Yes or No)	Formal Hypothesis Test Conclusion
1.	H_0	Low	Yes. Evidence	There is significant evidence to reject the claim.
2.	H_A	High		
3.	H_A	High	No. Not evidence.	There is not significant evidence to support the claim.
4.	H_0	Low		
5.	H_0	High	No. Not evidence.	There is not significant evidence to reject the claim.
6.	H_A	High		
7.	H_A	Low	Yes. Evidence	There is significant evidence to support the claim.



8.	H_0	High		
9.	H_0	Low	Yes. Evidence	There is significant evidence to reject the claim
10.	H_A	Low		
11.	H_A	High	No. Not evidence.	There is not significant evidence to support the claim.
12.	H_0	High		
13.	H_0	Low	Yes. Evidence	There is significant evidence to reject the claim.
14.	H_A	High		
15.	H_A	Low	Yes. Evidence	There is significant evidence to support the claim.
16.	H_0	Low		
17.	H_0	High	No. Not evidence.	There is not significant evidence to reject the claim.
18.	H_A	Low		
19.	H_A	High	No. Not evidence.	There is not significant evidence to support the claim.
20.	H_0	Low		
21.	H_0	Low	Yes. Evidence	There is significant evidence to reject the claim.
22.	H_A	High		
23.	H_A	High	No. Not evidence.	There is not significant evidence to support the claim.
24.	H_0	High		
25.	H_0	Low	Yes. Evidence	There is significant evidence to reject the claim.
26.	H_A	Low		
27.	H_A	High	No. Not evidence.	There is not significant evidence to support the claim.
28.	H_0	Low		
29.	H_0	High	No. Not evidence.	There is not significant evidence to reject the claim.
30.	H_A	Low		

31.

- a) Since the P-value is lower than the significance level, we should reject the null hypothesis.
- b) Since the P-value is lower than the significance level, we do have significant evidence.
- c) There is significant evidence to support the claim that less than 4% of people showed side effects to the medication.
- d) Statistical evidence agrees with the hospital that less than 4% of the population will show side effects to the medication.

33.

- a) Since the P-value is lower than the significance level, we should reject the null hypothesis.
- b) Since the P-value is lower than the significance level, we do have significant evidence.
- c) There is significant evidence to reject the claim that the candidate will receive 54% of the vote.
- d) Statistical evidence disagrees with the statement that the candidate will receive 54% of the vote.

35.

- a) Since the P-value is lower than the significance level, we should reject the null hypothesis.
- b) Since the P-value is lower than the significance level, we do have significant evidence.
- c) There is significant evidence to reject the claim that at least 50% of patients taking Toprol have seen improvement in their migraine symptoms.
- d) Statistical evidence disagrees with the statement that at least 50% of patients taking Toprol have seen improvement in their migraine symptoms. Statistical evidence suggest that it is more likely to be less than 50%.



37.

- a) Since the P-value is higher than the significance level, we should fail to reject the null hypothesis.
 - b) Since the P-value is higher than the significance level, we do not have significant evidence.
 - c) There is not significant evidence to support the claim that a higher percentage of people will vote for the democratic candidate over the republican candidate.
 - d) We do not have statistical evidence that agrees with the higher percentage for the democratic candidate. The voting percentages for the candidates seem very close. We do not have evidence one way or the other.
-

Section 3E Odd Answers

1.

Type 2 Error

3.

Beta Level

5.

Confidence Level

7.

Increase the sample size or increase the significance level.

9.

The probability of type 1 error (alpha level) will increase. The probability of type 2 error (beta level) will decrease.

11.

5% significance level (5% alpha level)

13.

A high P-value from biased data could result in a type 2 error.

15.

- a) A type 1 error would be that the company believes that very few airbags are defective when in reality many are defective. The company would decide to not recall the cars. This would be a grave mistake if many of the airbags are defective. The result would be deaths, injuries and lawsuits against the company for defective airbags.
- b) A type 2 error would be that the company believes that many airbags are defective when in reality hardly any are defective. The company would decide recall the cars by mistake. This would result in a loss of money for the company since mechanics would replace many airbags that did not need to be replaced. The company may also face a loss of reputation since the airbag recall may scare customers from purchasing their cars in the future.
- c) In this situation, the type 1 error was much more serious than the type 2 error. I would recommend the company avoid the type 1 error even if they make a type 2 error. So the company should decrease the significance level (alpha level) to 1% or even 0.5%. I would also recommend collecting more data before making a decision.



17.

a) A type 1 error would be Trisha believing that the stock price will drop when in reality it won't. She will sell the stock when the stock price will either remain the same or increase. Selling stock too early would result in a loss of money.

b) A type 2 error would be Trisha believing that the stock price will remain the same or increase when in reality it will decrease. She will hold onto the stock when the stock price will decrease. Not selling the stock when the price will decrease would result in a loss of money.

c) Both type 1 and type 2 errors seem like they result in a loss of money. I would recommend leaving the significance level (alpha level) at 5% since this balances the errors. Trisha could also collect more data before making a decision.

19.

a) A type 1 error would be team believes the players' scoring will decrease when in reality it will increase. The team would decide to not resign a player that would be a real difference maker on their team. The result would be that they save money, but the team winning percentage may drop dramatically. This may result in a loss of revenue for the team as fans may not support the team, not buy tickets, or not buy merchandise.

b) A type 2 error would be team believes the players' scoring will increase or stay the same when in reality it will dramatically decrease. The team would decide to resign a player for a large contract when the player is no longer worth the contract. The result would be a huge loss of money without seeing an increase in the team's overall winning percentage. This may result in fans not supporting the team, not buying tickets, or not buying merchandise.

c) Both errors seem to be bad. The type 1 error may be slightly better since the saving of money on the contract may offset partially the loss of revenue for the team not winning. Though they would probably just spend that money on another player that they have more confidence in. Both type 1 and type 2 errors seem like they result in a loss of money. I would recommend leaving the significance level (alpha level) at 5% since this balances the errors.

Section 3F Odd Answers

1.

One-population Proportion Assumptions

- The categorical sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least ten successes and at least ten failures.

2.

One-population Mean Assumptions

- The quantitative sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- The sample size should be at least 30 or have a nearly normal shape.

3.

One-Population Randomized Simulation Assumptions

- The sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.



5.

One-population Proportion Assumptions

The categorical sample data should be collected randomly or be representative of the population. No. This was not a random sample. It was convenience data and would have a large amount of bias.

Data values within the sample should be independent of each other. No. People in the same store may be related or friends.

There should be at least ten successes and at least ten failures. No. There was at least 10 failures (71), but there was not at least 10 successes (8).

This data does not pass all the assumptions and should not be used to test a claim about the population.

7.

One-population Proportion Assumptions

The categorical sample data should be collected randomly or be representative of the population. No. This was not a random sample. It was convenience data and would have a large amount of bias.

Data values within the sample should be independent of each other. No. People in the same English class will be related or friends.

There should be at least ten successes and at least ten failures. No. There was at least 10 successes (26), but there was not at least 10 failures (8).

This data does not pass all the assumptions and should not be used to test a claim about the population.

9.

One-population Mean Assumptions

The quantitative sample data should be collected randomly or be representative of the population. Yes the sample data was collected randomly.

Data values within the sample should be independent of each other. Yes. Since Jimmy took a small random sample from a large population of homes in Oklahoma city, the homes are likely independent of each other. It is unlikely they are on the same street or owned by the same owner.

The sample size should be at least 30 or have a nearly normal shape. Yes. Even though the sample size was below 30 (28) the histogram was normal (bell-shaped).

This data does pass all of the assumptions and can be used to test the claim about the population.

11.

One-population Mean Assumptions

The quantitative sample data should be collected randomly or be representative of the population. Yes. The data was collected randomly. A random cluster technique was used.

Data values within the sample should be independent of each other. Maybe not. Despite having a small random sample from a large population, all the data came from the same three streets. People living on the same street likely have similar socio-economic levels.

The sample size should be at least 30 or have a nearly normal shape. Yes. Despite the histogram being skewed right, the sample size was greater than 30 (63).

This data does not pass all the assumptions and should not be used to test a claim about the population.



13.

 $H_0: \pi = 0.13$ (claim) $H_A: \pi \neq 0.13$

Two-tailed test

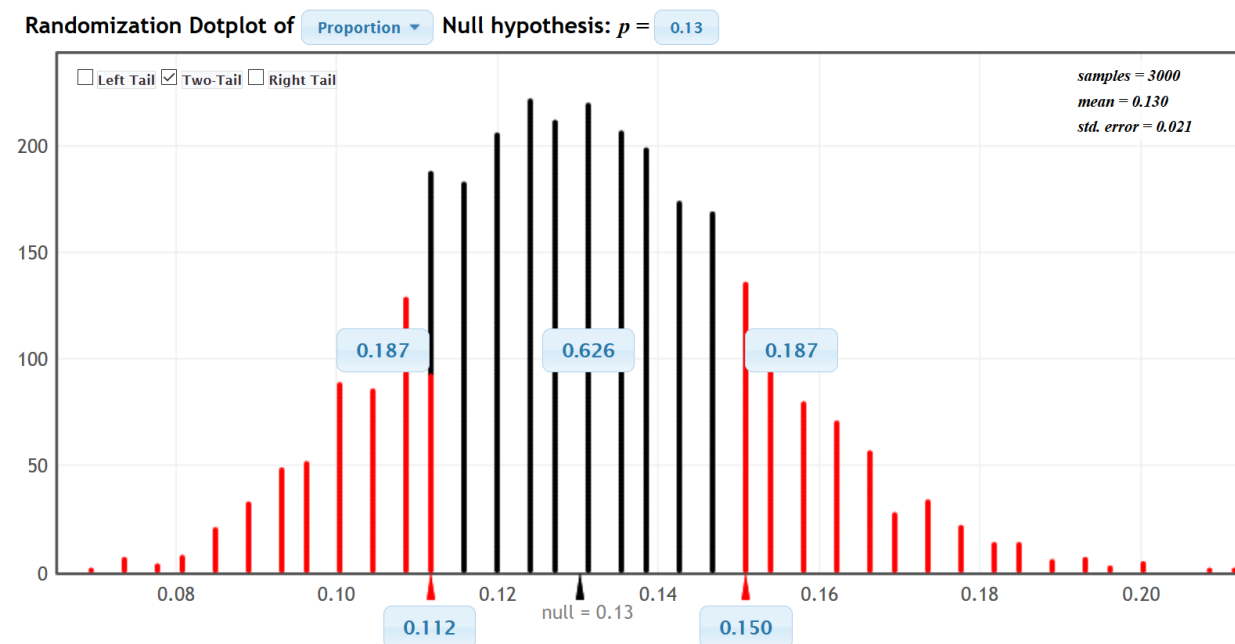
One-Population Randomized Simulation Assumptions

The sample data should be collected randomly or be representative of the population. **Yes.** The sample data was collected randomly.

Data values within the sample should be independent of each other. **Maybe.** Since it was a small random sample from a large population of various materials we are unlikely to get materials by the same author but we may get materials from the same publisher since there are not that many publishers. Publishers may have similar policies about using digital materials.

Assuming the materials are independent, we will proceed with the test.

The P-value and test statistic will vary slightly because of sampling variability.

P-value Simulation

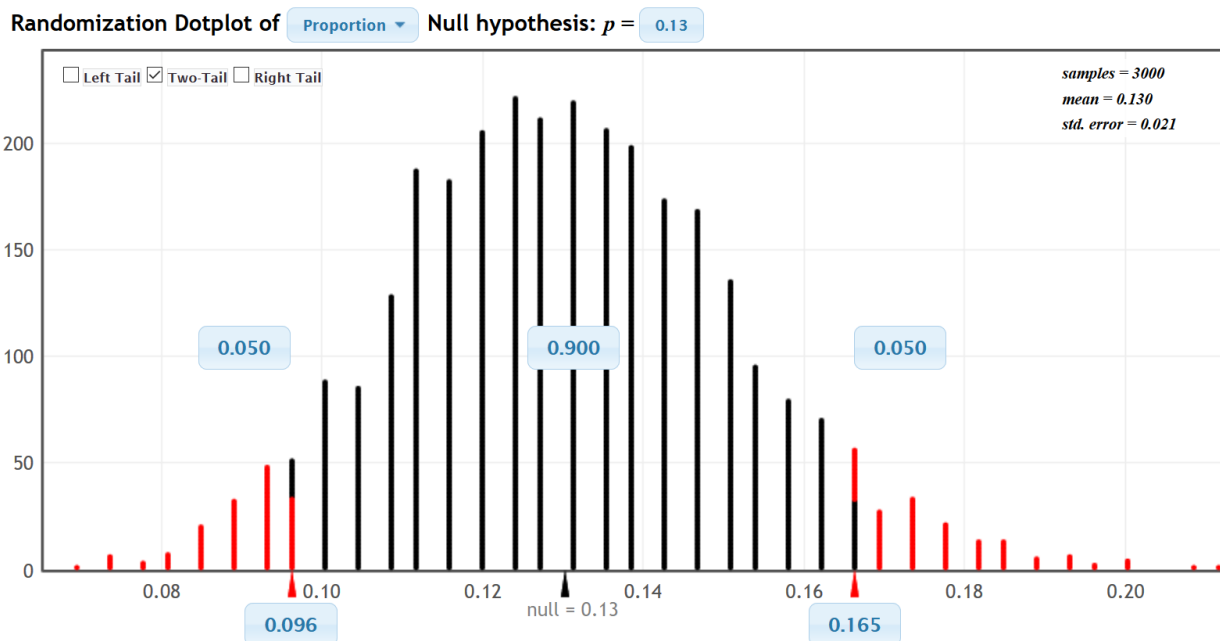
(#13 continued)

Approximate P-value = $0.187 + 0.187 = 0.374$ or 37.4%

P-value sentence: If the null hypothesis is true and 13% of materials are digital, then there is a 37.4% probability of getting the sample data or more extreme because of sampling variability.

Critical Value Simulation (Simulations will vary.)





Notice the sample proportion of (0.15) does not fall in the tail determined by the simulation and the 10% significance level (5% in each tail). The sample data does not significantly disagree with the null hypothesis.

Fail to reject the null hypothesis.

Conclusion: There is not significant evidence to reject the claim that 13% of materials are digital.

The population percentage could be 13% since random sample data does not significantly disagree with it. However we do not have evidence.

Approximate Z-test statistic (answers will vary) = $(0.15 - 0.13) \div 0.021 \approx +0.952$ standard errors. (Not significant.)

15.

$H_0: \pi = 0.2$

$H_A: \pi > 0.2$ (claim)

Right-tailed test

One-Population Randomized Simulation Assumptions

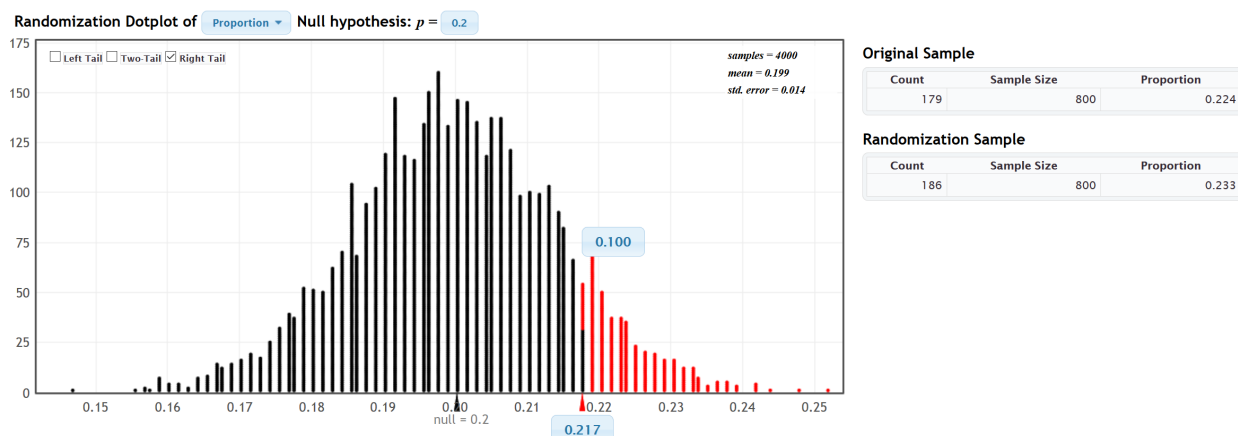
The sample data should be collected randomly or be representative of the population. **Yes.** The sample data was collected randomly.

Data values within the sample should be independent of each other. **Yes.** Since this is a small random sample of children from a large population, they are likely to be independent.

The P-value and test statistic will vary slightly because of sampling variability.

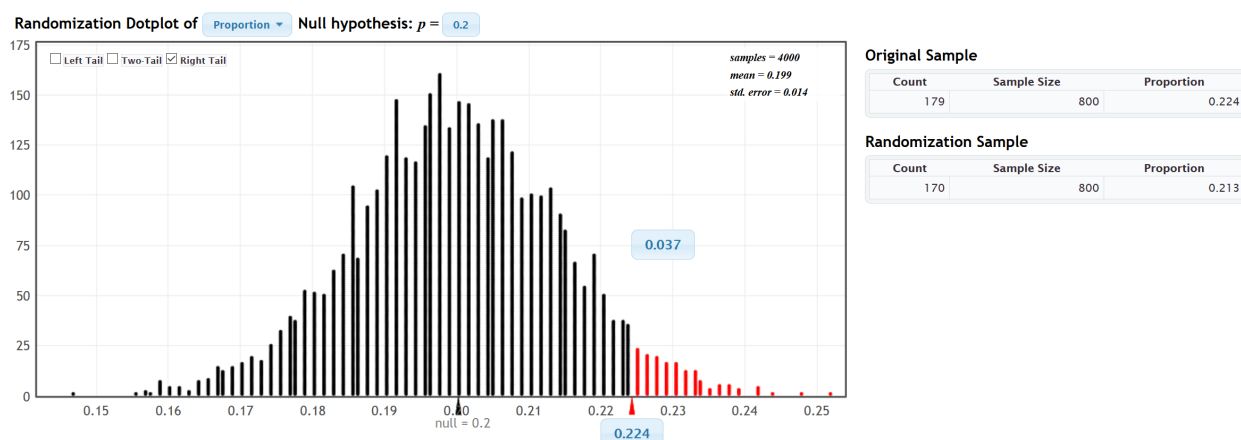
(#15 continued) Critical Value Simulation (Simulations will vary.)





Notice the sample proportion of (0.224) does fall in the right tail determined by the simulation and the 10% significance level. The sample data significantly disagrees with the null hypothesis.

P-value Simulation (will vary)



Approximate P-value = 0.037 or 3.7%

P-value sentence: If the null hypothesis is true and 20% of children are obese, then there is a 3.7% probability of getting the sample data or more extreme because of sampling variability.

Reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that more than 20% of children are obese.

We have statistical evidence that the population percentage of obese children is more than 20%

Approximate Z-test statistic (answers will vary) = $(0.224 - 0.2) \div 0.014 \approx +1.714$ standard errors. (Significant.)

17.

$H_0: \pi = 0.5$

$H_A: \pi > 0.5$ (claim)

Right-tailed test

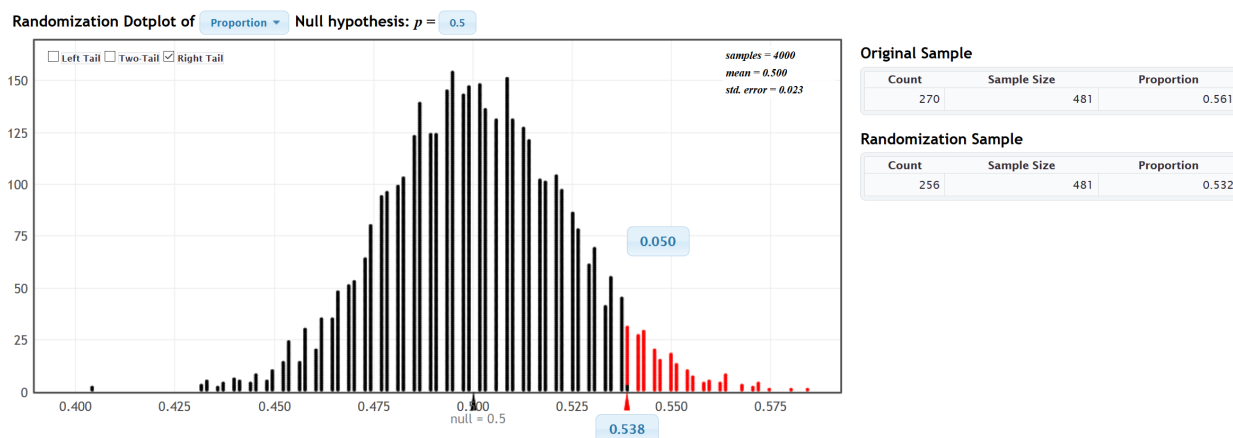
One-Population Randomized Simulation Assumptions

The sample data should be collected randomly or be representative of the population. Yes. The sample data was not collected randomly, however since it was a census of one semester, it may be representative.



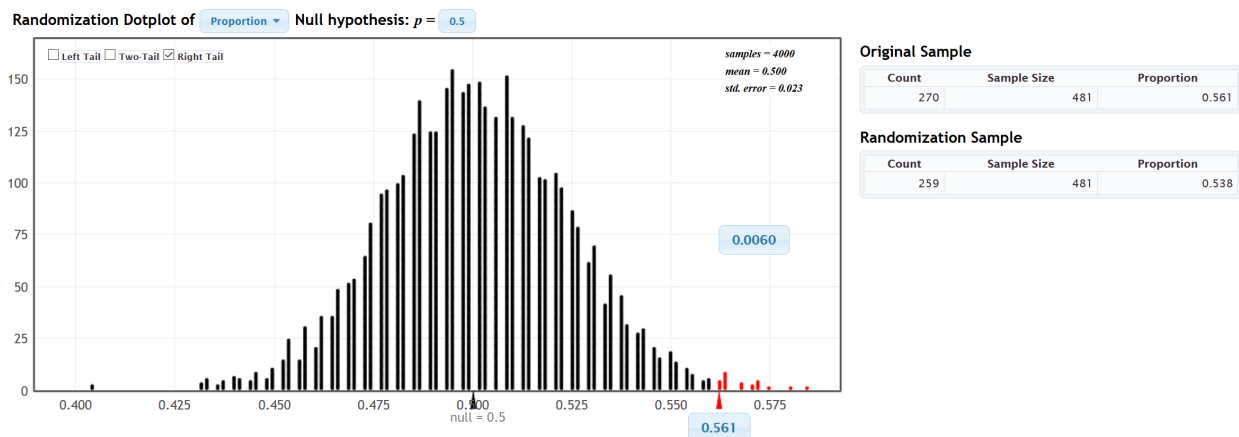
Data values within the sample should be independent of each other. **No.** The students came from the same statistics classes.

The P-value and test statistic will vary slightly because of sampling variability.



Notice the sample proportion of (0.561) falls in the right tail determined by the simulation and the 5% significance level. The sample data significantly disagrees with the null hypothesis.

P-value Simulation (will vary)



Approximate P-value = 0.006 or 0.6%

P-value sentence: If the null hypothesis is true and 50% of math 075 students are female, then there is a 0.6% probability of getting the sample data or more extreme because of sampling variability.

Reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that more than 50% of math 075 students are female.

Approximate Z-test statistic (answers will vary) = $(0.561 - 0.5) \div 0.023 \approx +2.65$ standard errors. (Significant.)

19.

$H_0 : \mu = 240$ feet

$H_A : \mu > 240$ feet (claim)

Assumptions

Random sample? Yes. Given it is a random sample.



Individual trees independent? Yes. A small random sample out of a large population. The trees are likely to be independent and not all measured from the same area.

At least 30 or normal? The shape is unknown so we must have a sample size of at least 30. There was 47 trees so it does pass the "30 or normal" requirement.

Z-test statistic = 2.109

The sample mean of 248 feet was 2.109 standard errors above the population mean of 240 feet. This indicates that the sample data does significantly disagree with the null hypothesis since the test statistics falls in the right tail corresponding to the critical value.

P-value = 0.0202 = 2.02%

If the null hypothesis is true and the population mean average height of redwood trees is 240 feet, there is a 2.02% probability of getting the sample data or more extreme because of sampling variability.

Since the P-value is less than our 5% significance level, it is unlikely that the sample data occurred by random chance (sampling variability).

Reject H_0

There is significant evidence to support the claim that the population mean average height of redwood trees is over 240 feet.

21.

$H_0 : \mu = \$3.50$

$H_A : \mu > \$3.50$ (claim)

Assumptions

Random sample? Yes. Given it is a random sample.

Individual independent? Yes. A small random sample out of a large population. The hamburgers probably did not come from the same restaurant.

At least 30 or normal? Even though the sample size is below 30 (24), the histogram looks normal. so it does pass the "30 or normal" requirement.

Z-test statistic = 1.633

The sample mean of \$3.88 was 1.633 standard errors above the population mean of \$3.50. This indicates that the sample data does significantly disagree with the null hypothesis since the test statistics falls in the right tail corresponding to the critical value.

P-value = 0.0508 = 5.08%

If the null hypothesis is true and the population mean average price of a hamburger is \$3.50, there is a 5.08% probability of getting the sample data or more extreme because of sampling variability.

Since the P-value is less than our 10% significance level, it is unlikely that the sample data occurred by random chance (sampling variability).

Reject H_0

There is significant evidence to support the claim that the population mean average price of a hamburger is greater than \$3.50.

23.

$H_0 : \mu = 160$ pounds

$H_A : \mu < 160$ pounds (claim)

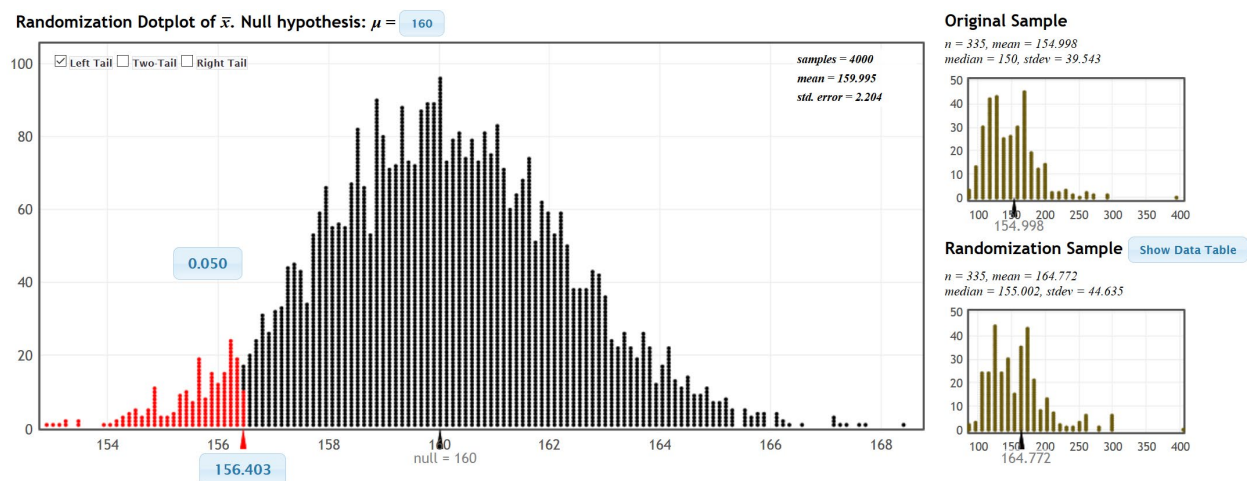


Assumptions for randomized simulation

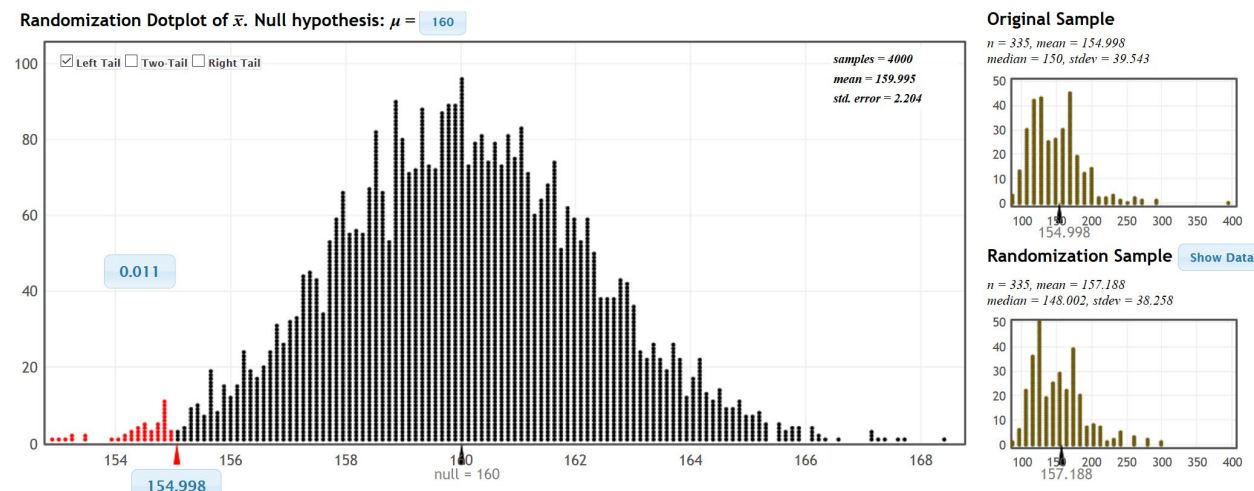
Random sample or representative? Yes. Even though the data was not a random sample, it was a census of all math 140 students in the semester. It is probably representative of all math 140 students from all semesters.

Individual independent? No. These students came from the same math 140 classes.

Randomized simulations and P-values will vary



The sample mean average weight was 154.998 pounds which did fall in the left tail determined by the simulation and the 5% significance level. This indicates that the sample data does significantly disagree with the null hypothesis.



The estimated P-value from this simulation was 0.011 or 1.1%. If the null hypothesis is true, there is a 1.1% probability of getting the sample data or more extreme by random chance.

Since the P-value is less than the significance level of 5%, it is unlikely this sample data occurred by random chance (sampling variability).

Reject H_0

There is significant evidence to support the claim that the population mean average weight of math 140 students is less than 160 pounds.



Chapter 3 Review All Answers

1.

Hypothesis Test: A procedure for testing a claim about a population.

Null Hypothesis (H_0): A statement about the population that involves equality. It is often a statement about "no change", "no relationship" or "no effect".

Alternative Hypothesis (H_A or H_1): A statement about the population that does not involve equality. It is often a statement about a "significant difference", "significant change", "relationship" or "effect".

Population Claim: What someone thinks is true about a population.

Test Statistic: A number calculated in order to determine if the sample data significantly disagrees with the null hypothesis. There are a variety of different test statistics depending on the type of data.

One-Population Proportion Test Statistic (z): The sample proportion is this many standard errors above or below the population proportion in the null hypothesis.

One-Population Mean Test Statistic (t): The sample mean is this many standard errors above or below the population mean in the null hypothesis.

Critical Value: A number we compare our test statistic to in order to determine significance. In a sampling distribution or a theoretical distribution approximating the sampling distribution, the critical value shows us where the tail or tails are. The test statistic must fall in the tail to be significant.

Sampling Variability: Also called "random chance". The principle that random samples from the same population will usually be different and give very different statistics. The random samples will usually be different than the population parameter.

P-value: The probability of getting the sample data or more extreme because of sampling variability (by random chance) if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. This is the probability of making a type 1 error. The P-value is compared to this number to determine significance and sampling variability. If the P-value is lower than the significance level, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened because of sampling variability.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value and standard error without a formula.

Type 1 Error: When biased sample data leads you to support the alternative hypothesis when the alternative hypothesis is actually wrong in the population.

Type 2 Error: When biased sample data leads you fail to reject the null hypothesis when the null hypothesis is actually wrong in the population.

Beta Level (β): The probability of making a type 2 error.

Conclusion: A final statement in a hypothesis test that addresses the claim and evidence.



2.

Randomized simulation is used to determine if sample data significantly disagrees with the null hypothesis and if the sample data occurred by random chance. The simulation can be used to calculate the P-value and determine the tail or tails and significance without a formula, test statistic, critical value, or theoretical curve. It also has less assumptions than traditional formula hypothesis tests.

3.

We can determine if sample data significantly disagrees with the null hypothesis in three ways.

If the test statistic falls in a tail determined by the critical value.

If the P-value is lower than the significance level.

If the sample statistic falls in a tail of the simulation determined by the significance level.

4.

Calculate the P-value. The P-value determines the probability of the sample data occurring by sampling variability if the null hypothesis was true.

5.

If the P-value is less than or equal to the significance level, we reject the null hypothesis.

If the P-value is greater than the significance level, we will fail to reject the null hypothesis.

6.

If the P-value is low, start the conclusion with “there is significant evidence”.

If the P-value is high, start the conclusion with “there is not significant evidence”.

If the claim is the null hypothesis, finish the conclusion with “to reject the claim”.

If the claim is the alternative hypothesis, finish the conclusion with “to support the claim”.

7.

One-population Mean Assumptions

- The quantitative sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- The sample size should be at least 30 or have a nearly normal shape.

One-population Proportion Assumptions

- The categorical sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.
- There should be at least ten successes and at least ten failures.

One-Population Randomized Simulation Assumptions

- The sample data should be collected randomly or be representative of the population.
- Data values within the sample should be independent of each other.

8. Fill out the following table regarding test statistics and critical values.

Test Statistic	Critical Value	Does sample significantly disagree with H_0 or not?
$T = +1.774$	± 2.751	Does not significantly disagree since test stat not in tail.
$Z = -2.481$	-1.96	Does significantly disagree since test stat in tail.
$T = -3.394$	± 2.566	Does significantly disagree since test stat in tail.
$Z = +1.362$	$+1.645$	Does not significantly disagree since test stat not in tail.



9. Fill out the following table regarding P-value and Significance levels.

P-value	P-value %	Significance Level	Sampling Variability or Unlikely	Reject H_0 or Fail to reject H_0 ?
0.0002	0.02%	5%	Unlikely since P-value low	Reject H_0
0.3327	33.27%	1%	Could be sampling variability since P-value High	Fail to reject H_0
1.84×10^{-5}	0.00184%	10%	Unlikely since P-value low	Reject H_0
0.0941	9.41%	5%	Could be sampling variability since P-value High	Fail to reject H_0

10. Fill out the following table to practice writing conclusions.

P-value	Claim	Write the Conclusion addressing Evidence and claim
Low	H_0	There is significant evidence to reject the claim.
High	H_A	There is not significant evidence to support the claim.
High	H_0	There is not significant evidence to reject the claim.
Low	H_A	There is significant evidence to support the claim.

11.

a.

$$H_0: \mu = 98.6^\circ\text{F}$$

$$H_A: \mu < 98.6^\circ\text{F} \text{ (Claim)}$$

Left-tailed test

b.

$$H_0: \pi_1 = \pi_2 \text{ (Claim)}$$

$$H_A: \pi_1 > \pi_2$$

Right-tailed test

c.

$$H_0: \mu_1 = \mu_2 \text{ (Claim)}$$

$$H_A: \mu_1 \neq \mu_2$$

Two-tailed test

12.

a.

A Type 1 Error occurs when biased sample data gives you a low P-value and leads you to reject the null hypothesis and support the alternative hypothesis, when the alternative hypothesis is actually wrong in the population.

b.

A Type 2 Error occurs when biased sample data gives you a high P-value and leads you to fail to reject the null hypothesis when the null hypothesis is actually wrong in the population.

c.

Alpha Level (α) or Significance Level

d.

Beta Level (β)



e.

Any time sample data does not reflect the population a type 1 or type 2 error may occur. It is often due to poor sampling techniques, not recognizing bias, or just sampling variability.

f.

To limit the chances of a type 1 error, decrease the significance level (alpha level).

g.

There are two ways to limit the chances of type 2 error. The preferred method is to raise the sample size (collect more data). If that is not possible, you can also raise the significance level.

h.

A 5% significance level tends to keep both type 1 and type 2 errors low.

i.

At a 1% significance level, there is a lower probability of type 1 error and a higher probability of type 2 error.

j.

At a 10% significance level, there is a higher probability of type 1 error and a lower probability of type 2 error.

13.

$H_0: \mu = 180$ pounds (claim)

$H_A: \mu \neq 180$ pounds

Two-tailed test

Assumptions Check

Random Sample? Yes. Given in the problem.

Individuals independent? Yes. This is a small random sample out of a huge population. The men are not likely to be related.

P-value (Two-tailed) = $0.030 + 0.030 = 0.060 = 6.0\%$

If the null hypothesis is true and the population mean average weight of all men is 180 pounds, then there is 6.0% probability of getting this sample data or more extreme by random chance.

Since the P-value is higher than our significance level, the sample data does not significantly disagree with the null hypothesis and could happen by random chance.

Fail to reject H_0 .

There is not significant evidence to reject the articles claim that the population mean average weight of all men is 180 pounds.

(The article could be correct. We do not have any evidence to contradict it.)

14.

$H_0: \mu = \$25$

$H_A: \mu > \$25$ (claim)

Right-tailed test

Assumptions Check



Random Sample? Yes. Given in the problem.

Individuals independent? Yes. This is a small random sample out of a huge population. The men are not likely to be related.

At least 30 or normal? The sample size is below 30, but since the histogram showed a normal shape, the data does pass the “30 or normal” requirement.

Test Statistic = 2.204

The sample mean of \$26.82 is 2.204 standard errors above the population mean of \$25. The test statistic indicates that the sample data significantly disagrees with the null hypothesis since it falls in the tail determined by the critical value.

P-value = 0.0181 = 1.81%

If the null hypothesis is true and the population mean average salary of nurses is \$25, then there is 1.81% probability of getting this sample data or more extreme by random chance.

Since the P-value is lower than our significance level, the sample data was unlikely to happen by random chance.

Reject H_0 .

There is significant evidence to support the claim that the population mean average salary for registered nurses is above \$25.

(The speaker at the convention is probably correct and we have evidence to back them up.)

15.

$H_0: \pi = 0.1$

$H_A: \pi > 0.1$ (claim)

Right-tailed test

Assumptions Check

Random Sample? Yes. Given in the problem.

Individuals independent? Yes. This is a small random sample out of a huge population. The women are not likely to be related.

P-value = 0.0015 = 0.15%

If the null hypothesis is true and the population proportion of women with a tattoo is 10%, then there is 0.15% probability of getting this sample data or more extreme by random chance.

Since the P-value is lower than our significance level, the sample data does significantly disagree with the null hypothesis and is unlikely to happen by random chance.

Reject H_0 .

There is significant evidence to support the claim that more than 10% of all women have at least one tattoo.

(The article is probably correct and we have evidence to back it up.)



Section 4A Odd Answers

	Type of Test	T-test stat	Sentence to explain T-test statistic.	Critical Value	Does the T-test statistic fall in a tail determined by a critical value? (Yes or No)	Are the sample means from the two groups significantly different or not? Explain.	Does sample data significantly disagree with H_0 ? Explain.
1.	Right Tailed	+1.383	The sample mean for group 1 was 1.383 standard errors above the sample mean for group 2	+2.447	No. The test statistic does not fall in the right tail.	No. The sample means are not significantly different since the test stat did not fall in the tail.	Sample data does not sig. disagree with H_0 since the test stat did not fall in the tail.
2.	Left Tailed	-2.851		-1.773			
3.	Two Tailed	-1.501	The sample mean for group 1 was 1.501 standard errors below the sample mean for group 2	± 2.006	No. The test statistic does not fall in either of the tails.	No. The sample means are not significantly different since the test stat did not fall in the tail.	Sample data does not sig. disagree with H_0 since the test stat did not fall in the tail.
4.	Right Tailed	+3.561		+1.692			
5.	Two Tailed	+0.887	The sample mean for group 1 was 0.887 standard errors above the sample mean for group 2	± 1.943	No. The test statistic does not fall in one of the tails.	No. The sample means are not significantly different since the test stat did not fall in the tail.	Sample data does not sig. disagree with H_0 since the test stat did not fall in the tail.
6.	Left Tailed	-1.003		-2.759			
7.	Two Tailed	-4.416	The sample mean for group 1 was 4.416 standard errors below the sample mean for group 2	± 1.994	Yes. The test statistic does fall in one of the tails	Yes. The sample means are significantly different since the test stat fell in the tail.	Sample data significantly disagrees with H_0 since the test stat fell in the tail.
8.	Right Tailed	+0.275		+1.839			
9.	Left Tailed	-1.461	The sample mean for group 1 was 1.461 standard errors below the sample mean for group 2	-1.674	No. The test statistic does not fall in the left tail.	No. The sample means are not significantly different since the test stat did not fall in the tail.	Sample data does not sig. disagree with H_0 since the test stat did not fall in the tail.
10.	Two Tailed	+2.330		± 2.138			



	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.0007	0.07%	If the null hypothesis is true, there is a 0.07% probability of getting the sample data or more extreme by random chance.	10%	0.10	Sample data unlikely to occur by random chance.	Reject H_0
12.	0.421			1%			
13.	8.71×10^{-5}	0.00871%	If the null hypothesis is true, there is a 0.00871% probability of getting the sample data or more extreme by random chance.	5%	0.05	Sample data unlikely to occur by random chance.	Reject H_0
14.	0.339			1%			
15.	0.076	7.6%	If the null hypothesis is true, there is a 7.6% probability of getting the sample data or more extreme by random chance.	5%	0.05	Sample data could occur by random chance.	Fail to reject H_0
16.	0			10%			
17.	0.528	52.8%	If the null hypothesis is true, there is a 52.8% probability of getting the sample data or more extreme by random chance.	5%	0.05	Sample data could occur by random chance.	Fail to reject H_0
18.	0.0277			10%			
19.	3.04×10^{-6}	0.000304%	If the null hypothesis is true, there is a 0.000304% probability of getting the sample data or more extreme by random chance.	1%	0.01	Sample data unlikely to occur by random chance.	Reject H_0
20.	0.178			5%			

21.

Two quantitative data sets are considered matched pair if there is a one-to-one pairing between the data values. The first number in the first data set is directly related to the first number in the second data set. The second number in the first data set is directly related to the second number in the second data set, and so on. It is usually the same person measured twice.

Two quantitative data sets are considered independent groups if there is not a one-to-one pairing and individuals between the groups are not related. This is usually the case when we are comparing two separate groups like a random sample of adults to a random sample of children.



23.

Two-population Mean Assumptions (Not Matched Pair, Independent groups)

- The two quantitative samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

25.

Two-population Mean Assumptions (Matched Pair for Experiments. Same people or objects measured twice.)

- Quantitative ordered pair data
- Data values within the samples should be independent of each other.
- There should be at least thirty ordered pairs or the differences should have a nearly normal shape.

Two-population Mean Assumptions (Independent groups for Experiment)

- The two quantitative samples should be randomly assigned from the people or objects in the experiment.
- Data values within each sample should be independent of each other.
- Data values between the two samples should be independent of each other.
- The sample sizes should be at least 30 or have a nearly normal shape.

27.

To prove cause and effect we will need to set up a controlled experiment and control confounding variables. It must meet the assumptions for a two-population mean experiment. If the T-test statistic falls in the tail determined by the critical value and the P-value is lower than the significance level, then this would prove cause and effect.

29.

These are independent groups and not matched pair.

μ_1 : Population mean average weight of male German Shepherds

μ_2 : Population mean average weight of male Dobermans

$H_0 : \mu_1 = \mu_2$ (The breed of dog is not related to the weight.)

$H_a : \mu_1 > \mu_2$ (claim) (The breed of dog is related to the weight.)

Right-tailed test

Assumptions Check

Both random samples? Yes. Both samples were collected randomly.

Individual dogs within the sample and between the samples should be independent? Yes. A small random sample out of large population will probably not have dogs that are related, owned by the same owner, or have the exact same diet.

At least 30 or normal? Both sample sizes were less than 30 (20 and 14), but since both samples were normal, it does pass the 30 or normal requirement.

T-test Statistic = 0.558

The sample mean weight for the German Shepherds was 0.558 standard errors above the sample mean weight for the Dobermans.

Since the test statistic did not fall in the right tail determined by the critical value, the sample mean for the German Shepherds was not significantly different than the sample mean for the Dobermans.

P-value = 0.2906 = 29.06%



If the null hypothesis is true, there is a 29.06% probability of getting this sample data or more extreme by random chance.

Since the p-value is higher than the significance level, this sample data could have occurred by random chance.

Conclusion

There is not significant evidence to support the claim that the population mean average weight of male German Shepherds is significantly higher than the population mean average weight of Dobermans.

Relationship

The weights were not significantly different. This indicates that being a German Shepherd or Doberman is not related to the weight.

31.

These are independent groups and not matched pair.

μ_1 : Population mean average gas mileage (mpg) for cars made in the U.S.

μ_2 : Population mean average gas mileage (mpg) for cars not made in the U.S.

$H_0 : \mu_1 = \mu_2$ (The country a car is made in is not related to the gas mileage.)

$H_a : \mu_1 < \mu_2$ (claim) (The country a car is made in is related to the gas mileage.)

Left-tailed test

Assumptions Check

Both random samples? Yes. Both samples were collected randomly.

Individual cars within the sample and between the samples should be independent? Probably not. The cars may have been made by the same manufacturer. Also the sample size is not significantly smaller than the population size.

At least 30 or normal? No. Both sample sizes were below 30 and not normal. The U.S. cars looked skewed right and the cars outside the U.S. data looked bimodal.

T-test Statistic = -2.001

The sample mean mpg for U.S. cars was 2.001 standard errors below the sample mean for cars outside the U.S.

Since the test statistic did fall in the left tail determined by the critical value, the sample mean mpg for the U.S. cars was significantly lower than the sample mean mpg for cars made outside the U.S.

P-value = 0.0273 = 2.73%

If the null hypothesis is true, there is a 2.73% probability of getting this sample data or more extreme by random chance.

Since the p-value is lower than the significance level, this sample data is unlikely to have occurred by random chance.

Conclusion

There is significant evidence to support the claim that the population mean average mpg for cars made in the U.S. is significantly lower than the population mean average mpg for cars made outside the U.S. This conclusion may not be true since the data did not pass all of the assumptions for inference.

Relationship

The weights were significantly different. This indicates that the country a car is made in may be related to the gas mileage (mpg).



33.

These are independent groups and not matched pair.

μ_1 : Population mean average horsepower for cars made in the U.S.

μ_2 : Population mean average horsepower for cars not made in the U.S.

$H_0 : \mu_1 = \mu_2$ (The country a car is made in is not related to the horsepower.)

$H_0 : \mu_1 > \mu_2$ (claim) (The country a car is made in is related to the horsepower.)

Right-tailed test

Assumptions Check

Both random samples? Yes. Both samples were collected randomly.

Individual cars within the sample and between the samples should be independent? Probably not. The cars may have been made by the same manufacturer. Also the sample size is not significantly smaller than the population size.

At least 30 or normal? No. Both sample sizes were below 30 and skewed right.

T-test Statistic = +2.526

The sample mean average horsepower for the U.S. cars was 2.526 standard errors above the sample mean average horsepower for the cars outside the U.S.

Since the test statistic did fall in the right tail determined by the critical value, the sample mean horsepower for the U.S. cars was significantly higher than the sample mean horsepower for cars made outside the U.S.

P-value = 0.0081 = 0.81%

If the null hypothesis is true, there is a 0.81% probability of getting this sample data or more extreme by random chance.

Since the p-value is lower than the significance level, this sample data is unlikely to have occurred by random chance.

Conclusion

There is significant evidence to support the claim that the population mean average horsepower for cars made in the U.S. is significantly higher than the population mean average horsepower for cars made outside the U.S. This conclusion may not be true since the data did not pass all of the assumptions for inference.

Relationship

The amount of horsepower were significantly different. This indicates that the country a car is made in may be related to the horsepower.

35.

Since these are the same people measured twice, this data is matched pair.

Since this is a controlled experiment we can move beyond a relationship and discuss cause and effect.

μ_d : The population mean average of the differences (Quiz pulse rate – Lecture pulse rate)

Population 1: Pulse rate of students taking a quiz.

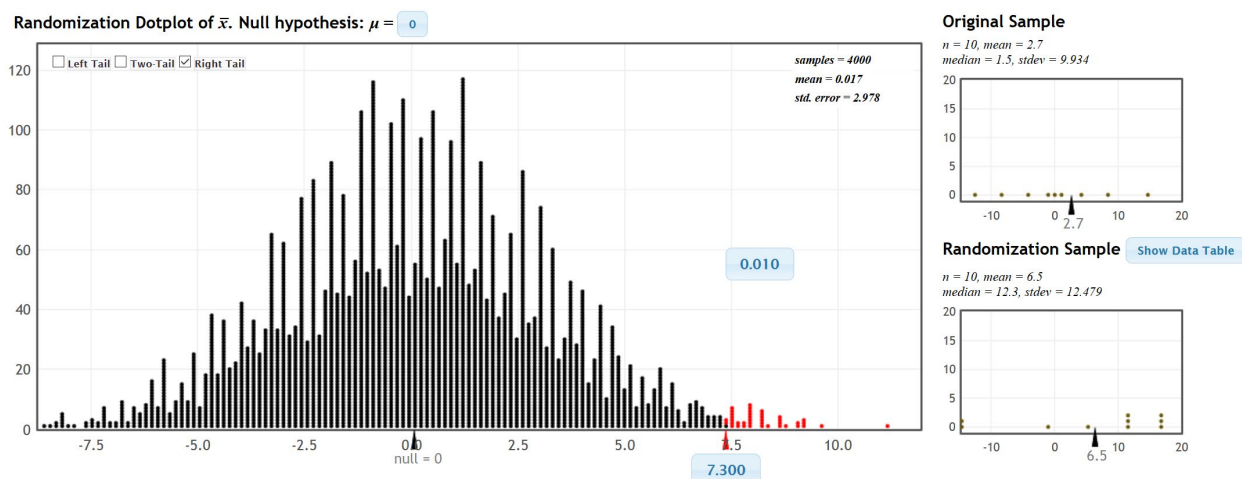
Population 2: Pulse rate of students attending lecture.

$H_0 : \mu_d = 0$ (Taking a quiz or attending lecture does not effect a persons' heartrate.)



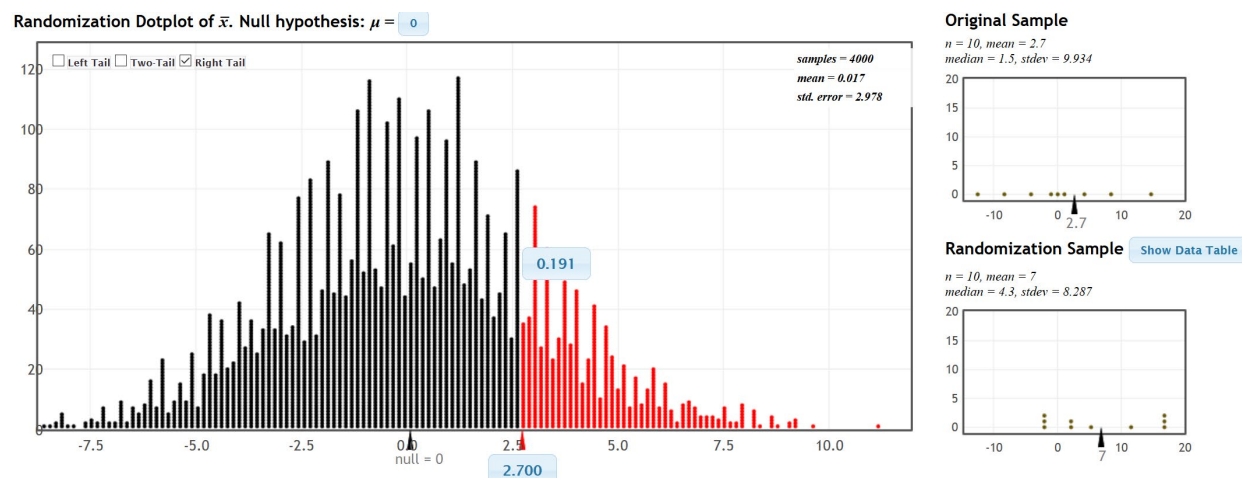
$H_0: \mu_d > 0$ (claim) (Taking a quiz or attending lecture does effect a persons' heartrate.)

Tail determined by the simulation and significance level (tails between simulations will vary)



The original sample mean does not fall in the tail determined by the simulation and significance level. The sample data does not significantly disagree with the null hypothesis.

P-value (P-values will vary)



This simulation indicates that the P-value is approximately 0.191 or 19.1%.

If the null hypothesis is true, then there is a 19.1% probability of getting the sample data or more extreme by random chance.

Since the P-value is higher than the significance level, the sample data could have occurred by random chance.

Fail to reject the null hypothesis.

Conclusion

There is not significant evidence to support the claim that heart rates when taking a quiz are significantly higher than when attending a lecture.

Cause and Effect

Since there was no significant difference between the heart rates on a quiz or lecture day, this indicates that taking a quiz or attending a lecture does not effect a persons' heart rate significantly.



37.

These are separate independent groups and are not matched pair.

Since the significance level was not given, we will use a 5% significance level.

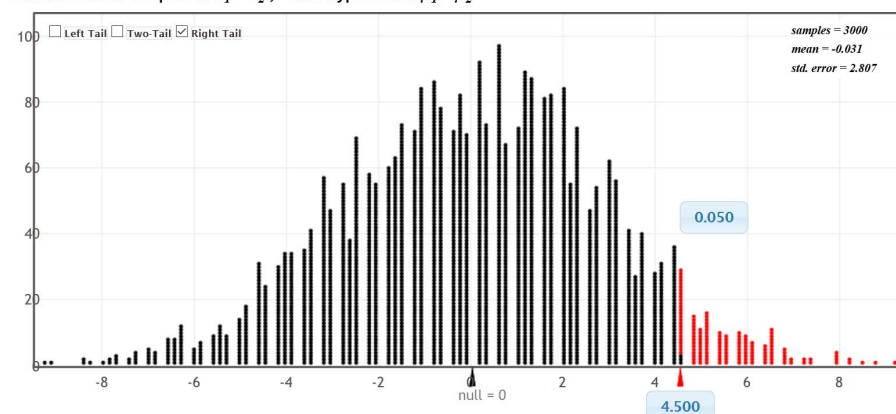
$H_0 : \mu_1 = \mu_2$ (Gender is not related to heart rate.)

$H_a : \mu_1 > \mu_2$ (claim) (Gender is related to heart rate.)

Right-tailed test

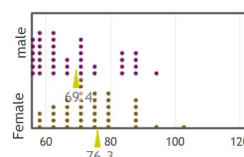
Simulation (Tail determined by the significance level.) (Simulations and tails will vary)

Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$



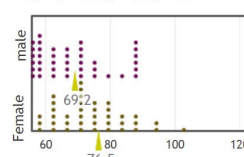
Original Sample

$\bar{x}_1 - \bar{x}_2 = 6.9$, $n_1 = 40$, $n_2 = 40$



Randomization Sample

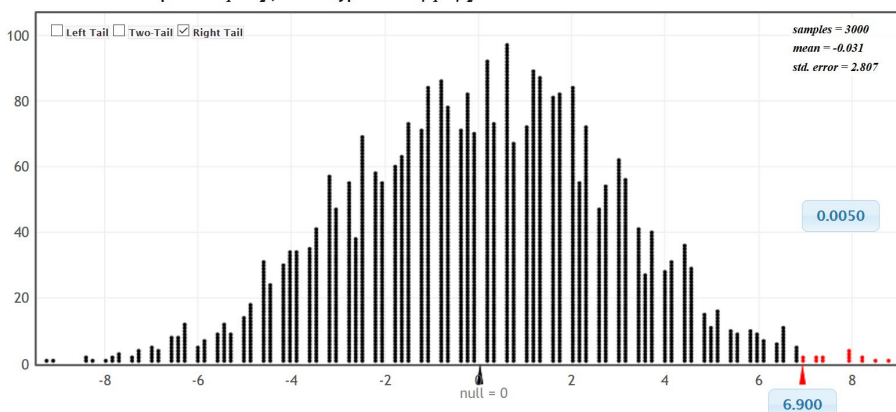
$\bar{x}_1 - \bar{x}_2 = 7.3$, $n_1 = 40$, $n_2 = 40$



Notice that the difference between the sample means does fall in the right tail determined by the significance level. The sample data does significantly disagree with the null hypothesis.

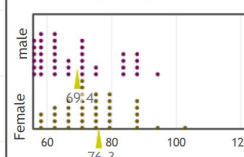
P-value (P-values from simulations will vary)

Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$



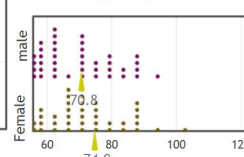
Original Sample

$\bar{x}_1 - \bar{x}_2 = 6.9$, $n_1 = 40$, $n_2 = 40$



Randomization Sample

$\bar{x}_1 - \bar{x}_2 = 4.1$, $n_1 = 40$, $n_2 = 40$



In this simulation, the P-value was 0.005 or 0.5%.

If the null hypothesis is true, there is a 0.5% probability of getting the sample data or more extreme by random chance.

Since the P-value is less than the significance level (5%), the sample data was unlikely to have occurred by random chance.

Reject H_0 .



Conclusion

There is significant evidence to support the claim that the population mean average heart rate for women is greater than the population mean average heart rate for men.

Relationship

The heart rates for the women and men were significantly different. This indicates that gender may be related to a persons' heart rate.

Section 4B Odd Answers

	F-test stat	Sentence to explain F-test statistic.	Critical Value	Does the F-test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	+5.573	The ratio of the variance between the groups to the variance within the groups is 5.573	+2.886	Yes. In Tail	Yes. Sample data significantly disagrees with H_0
2.	+1.192		+3.113		
3.	+0.664	The ratio of the variance between the groups to the variance within the groups is 0.664	+2.949	No. Not in tail.	No Sample data does not significantly disagree with H_0
4.	+4.415		+3.125		
5.	+3.718	The ratio of the variance between the groups to the variance within the groups is 3.718	+4.117	No. Not in tail.	No Sample data does not significantly disagree with H_0
6.	+0.991		+2.009		
7.	+2.652	The ratio of the variance between the groups to the variance within the groups is 2.652	+1.875	Yes. In Tail	Yes. Sample data significantly disagrees with H_0
8.	+1.585		+3.225		
9.	+2.447	The ratio of the variance between the groups to the variance within the groups is 2.447	+2.798	No. Not in tail.	No Sample data does not significantly disagree with H_0
10.	+8.133		+2.891		

	P-value Proportion	P-value %	Sentence to explain the P-value	Sig Level %	Sig level Prop	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.186	18.6%	If H_0 is true there is a 18.6% probability of getting the sample data or more extreme by random chance.	10%	0.10	Could be random chance	Fail to reject H_0
12.	0.0042			1%			



13.	2.59×10^{-4}	0.0259%	If H_0 is true there is a 0.0259% probability of getting the sample data or more extreme by random chance.	5%	0.05	Unlikely to be random chance	Reject H_0
14.	0.006			1%			
15.	0.353	35.3%	If H_0 is true there is a 35.3% probability of getting the sample data or more extreme by random chance.	5%	0.05	Could be random chance	Fail to reject H_0
16.	0			10%			
17.	0.041	4.1%	If H_0 is true there is a 4.1% probability of getting the sample data or more extreme by random chance.	5%	0.05	Unlikely to be random chance	Reject H_0
18.	0.274			10%			
19.	1.04×10^{-8}	0.00000104%	If H_0 is true there is a 0.00000104% probability of getting the sample data or more extreme by random chance.	1%	0.01	Unlikely to be random chance	Reject H_0
20.	0.067			5%			

21. The variance between the groups is calculated by taking the sample mean from each group and subtracting the mean of all the numbers in all the groups. They then square the differences and add up the sum of squares. The sum of squares between is then divided by the degrees of freedom between to get the variance between.

The variance within the groups takes each individual number in each data set and subtracts it from the sample mean of that group. They then square the differences and add up the sum of squares. The sum of squares within is then divided by the degrees of freedom within to get the variance within.

The F-test statistic is calculated by taking the variance between the groups and dividing by the variance within the groups.

23.

If the variance between the groups were about the same as the variance within the ratio between the variances would be close to one. This is a small F-test statistic and would indicate that the sample data does not significantly disagree with the null hypothesis.

25.

$H_0 : \mu_1 = \mu_2 = \mu_3$ (season is not related to the weight of bears)

H_A : at least one is \neq (season is related to the weight of bears) CLAIM

Assumptions

Random? Yes. Given in the problem.

Individuals between the groups and within the groups are independent? Probably Yes. As long as the bears were taken from different areas and were not the same bear measured multiple times.

All sample sizes at least 30 or normal? Yes. All the sample sizes were below 30 (13,24,17) but the histograms all looked normal.

Standard Deviations close? No. The standard deviation for the summer bears (22.463) is more than twice as large as for the spring bears (48.017).

F-test statistic = 13.55345



The ratio of the variance between the groups to the variance within the groups is 13.55345. This also indicates that the variance between the groups is 13.55345 times larger than the variance within the groups.

Since the F-test statistic falls in the tail determined by the critical value 5.0472, the sample data does significantly disagree with the null hypothesis. This also implies that the variance between the groups (22769.64632) is significantly higher than the variance within the groups (1679.98954).

P-value = 0.00002 = 0.002%

If the null hypothesis is true and the season is not related to the weights of the bears, then there is a 0.002% probability of getting this sample data or more extreme by random chance.

Since the P-value is lower than the significance level, the sample data was unlikely to have occurred because of sampling variability (random chance).

Reject the null hypothesis

Conclusion: There is significant evidence to support the claim that the season is related to the weight of bears. This conclusion is in question since the sample data did not pass all of the assumptions for inference.

The sample data indicates that the categorical variable (season) is probably related to the quantitative variable (weight). However, the data did not pass all of the assumptions for inference.

27.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (Political viewpoint is not related to the amount of alcohol consumed) CLAIM

H_A : at least one is \neq (Political viewpoint is not related to the amount of alcohol consumed)

Assumptions

Random or Representative? Yes. Even though the data was not random, a census of one semester of students may be representative of all pre-stat students from all semesters.

Individuals between the groups and within the groups are independent? Probably not. Students came from the same Math 075 classes and may be related.

All sample sizes at least 30 or normal? Yes. All of the histograms looked skewed right and were not normal. However, all of the sample sizes were above 30 (198, 111, 102, 90).

Standard Deviations close? Yes. None of the standard deviations were more than twice as large as any other.

F-test statistic = 0.89597

The ratio of the variance between the groups to the variance within the groups is 0.89597. This also indicates that the variance between the groups is very close to the variance within the groups.

Since the F-test statistic did NOT fall in the tail determined by the critical value 2.6228, the sample data does NOT significantly disagree with the null hypothesis. This also implies that the variance between the groups (8.48046) is NOT significantly higher than the variance within the groups (9.46512).

P-value = 0.44306 = 44.306%

If the null hypothesis is true and political views are not related to consuming alcohol, then there is a 44.306% probability of getting this sample data or more extreme by random chance.

Since the P-value is higher than the significance level, the sample data could have occurred simply because of sampling variability (random chance).

Fail to reject the null hypothesis



Conclusion: There is NOT significant evidence to reject the claim that political views are not related to consuming alcohol. This conclusion is in question since the sample data did not pass all of the assumptions for inference.

The sample data indicates that the categorical variable (political view) is probably not related to the quantitative variable (alcohol). However, the data did not pass all of the assumptions for inference.

29.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ (Country is NOT related to mpg)

H_A : at least one is \neq (Country is related to mpg) CLAIM

Original Sample

ANOVA Table

$n = 38, F = 3.406$

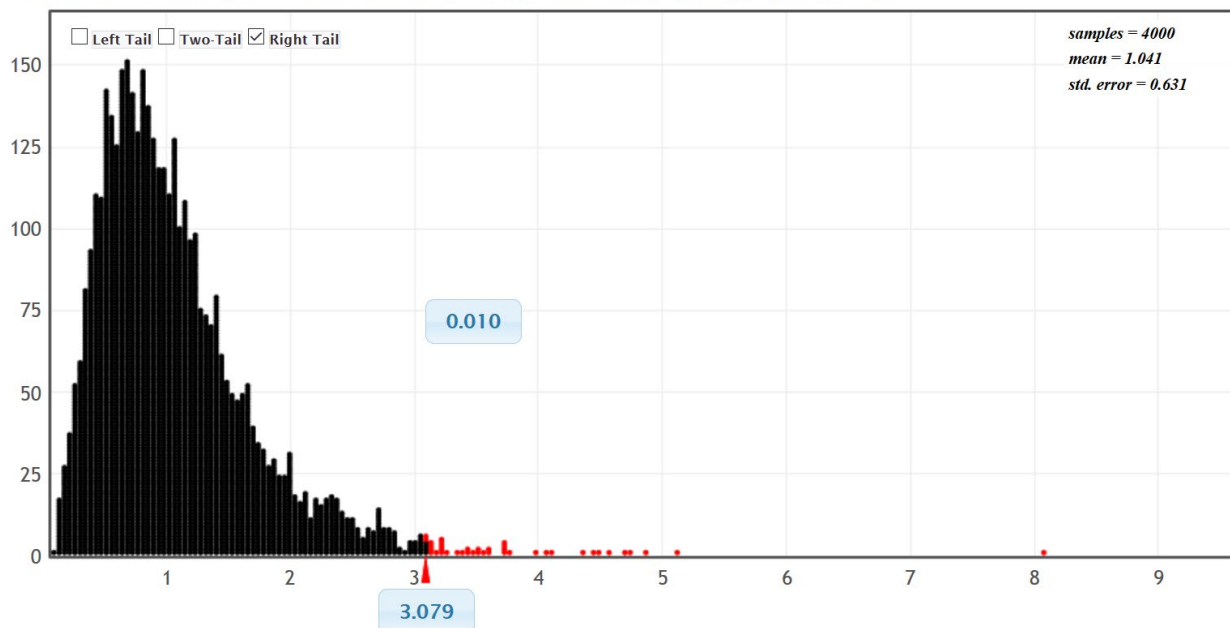
Statistics	U.S.	Japan	Germany	Sweden	France	Italy	Overall
Sample Size	22	7	5	2	1	1	38
Mean	23.0	29.6	27.1	19.3	16.2	37.3	24.8
Standard Deviation	6.1	4.5	5.7	3.3	NaN	NaN	6.5

F-test Statistic = 3.406

The ratio of the variance between the groups to the variance within the groups is 3.406.

Critical Value Calculation

Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

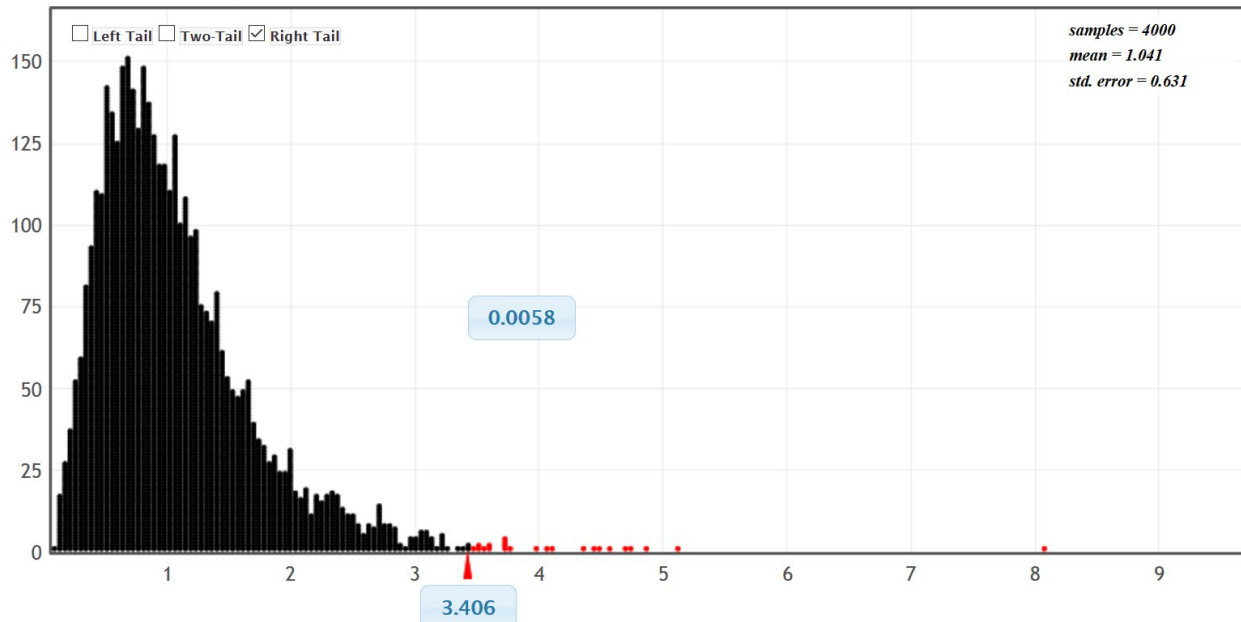


Critical Value = 3.079 (Answers were vary)

Since the test statistic 3.406 falls in the tail determined by the critical value 3.079, the sample data significantly disagrees with the null hypothesis.



P-value Calculation

Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ 

P-value = 0.0058 or 0.58% (Answers will vary)

Since the p-value was less than our significance level, it is unlikely for our sample data to have occurred because of sampling variability.

Since the p-value was less than our significance level, we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that the country a car is made in is related to the cars' gas mileage.

ANOVA Table				
	df	SS	MS	F
Groups	5	550.9	110.19	3.406
Error	32	1035.1	32.35	
Total	37	1586.1		

The variance between the groups (110.19) is significantly greater than the variance within the groups (32.35). We know it is significant since our F-test statistic fell in the tail determined by the critical value.

The P-value was lower than the significance level. This indicates that the categorical variable (country) is related to the quantitative variable (mpg).

31.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ (Country is NOT related to horsepower) CLAIM

H_A : at least one is \neq (Country is related to horsepower)



Original Sample ANOVA Table

$n = 38, F = 3.099$

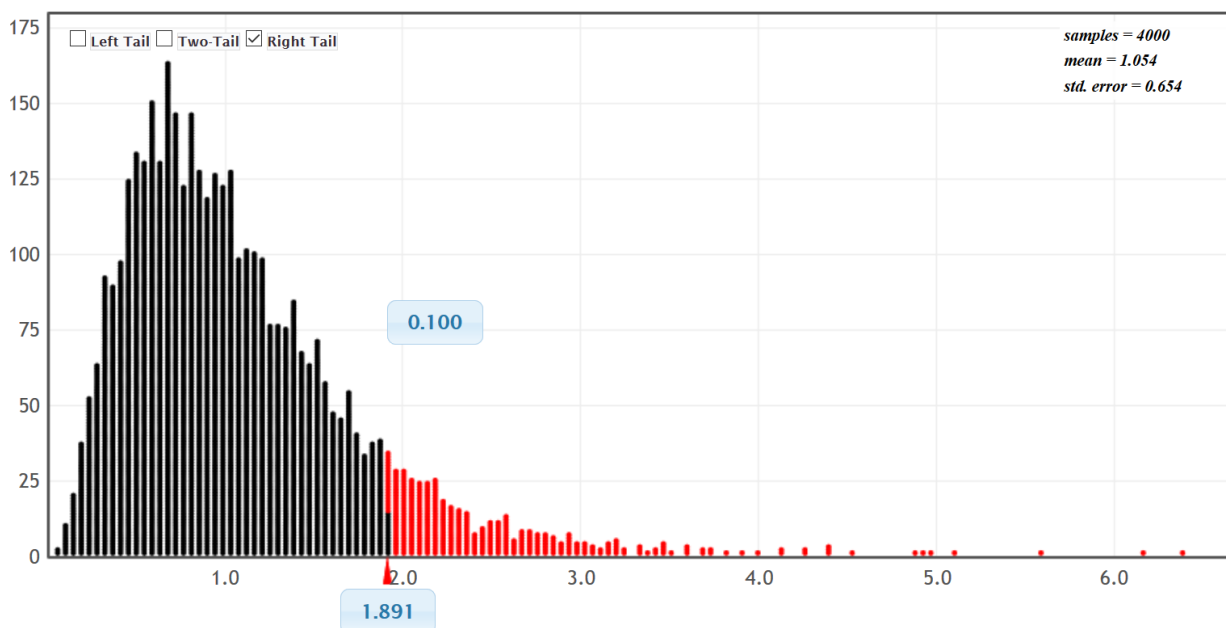
Statistics	U.S.	Japan	Germany	Sweden	France	Italy	Overall
Sample Size	22	7	5	2	1	1	38
Mean	110.2	81.0	86.6	120.0	133.0	69.0	101.7
Standard Deviation	26.4	15.2	18.6	7.1	NaN	NaN	26.4

F-test Statistic = 3.099

The ratio of the variance between the groups to the variance within the groups is 3.099.

Critical Value Calculation

Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

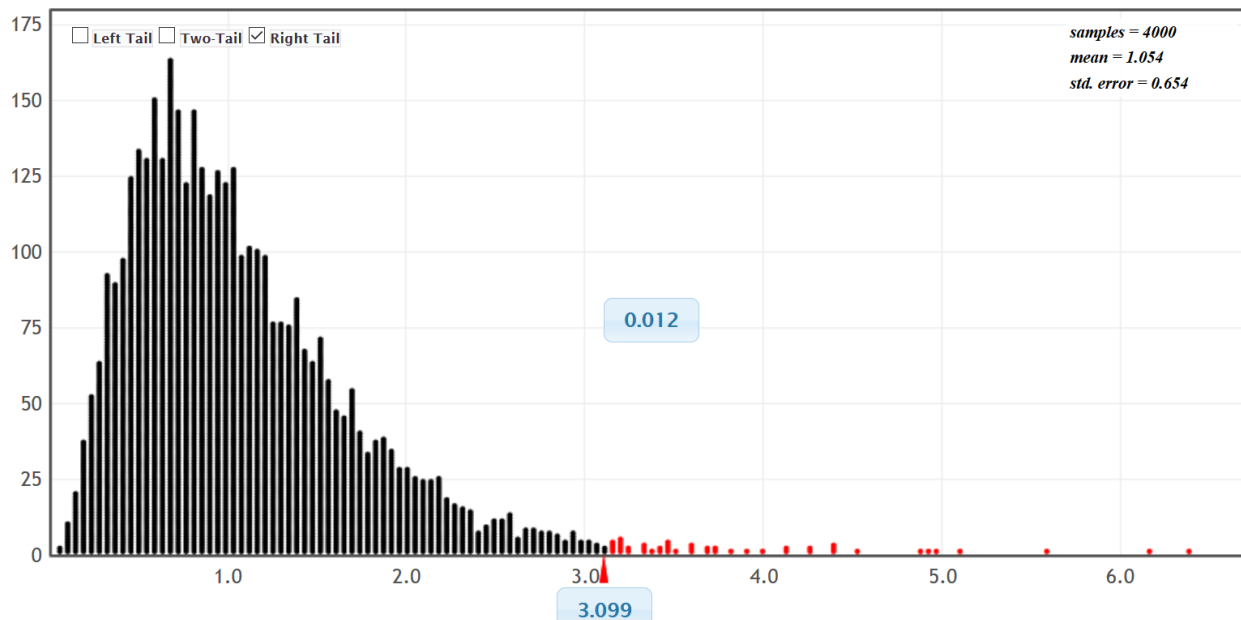


Critical Value = 1.891 (Answers were vary)

Since the test statistic 3.099 falls in the tail determined by the critical value 1.891, the sample data significantly disagrees with the null hypothesis.



P-value Calculation

Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ 

P-value = 0.012 or 1.2% (Answers will vary)

Since the p-value was less than our significance level, it is unlikely for our sample data to have occurred because of sampling variability.

Since the p-value was less than our significance level, we will reject the null hypothesis.

Conclusion: There is significant evidence to reject the claim that the country a car is made in is not related to the cars' horsepower.

ANOVA Table				
	df	SS	MS	F
Groups	5	8440.9	1688.2	3.099
Error	32	17434.5	544.8	
Total	37	25875.4		

The variance between the groups (1688.2) is significantly greater than the variance within the groups (544.8). We know it is significant since our F-test statistic fell in the tail determined by the critical value.

The P-value was lower than the significance level. This indicates that the categorical variable (country) is related to the quantitative variable (horsepower).

33.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (State is NOT related to price of home)

$H_A : \text{at least one is } \neq$ (State is related to price of home) CLAIM



Original Sample

ANOVA Table

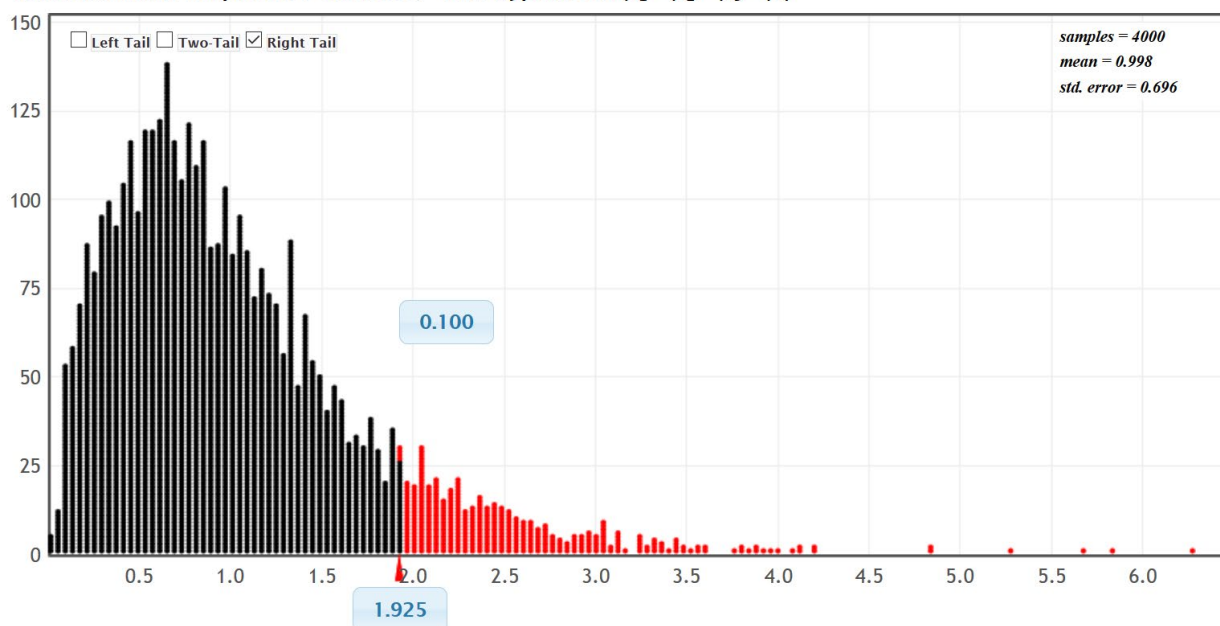
 $n = 120, F = 2.747$

Statistics	NJ	NY	PA	CA	Overall
Sample Size	30	30	30	30	120
Mean	388.5	565.6	249.6	715.1	479.7
Standard Deviation	224.7	697.6	179.3	1112.2	686.6

F-test Statistic = 2.747

The ratio of the variance between the groups to the variance within the groups is 2.747.

Critical Value Calculation

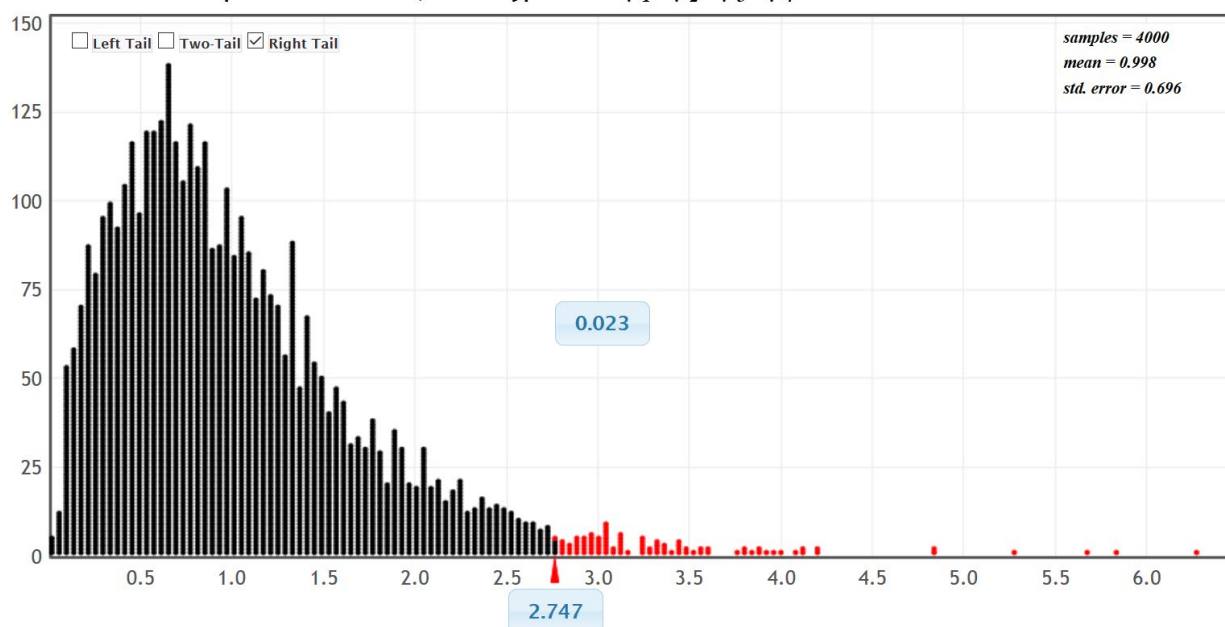
Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ 

Critical Value = 1.925 (Answers were vary)

Since the test statistic 2.747 falls in the tail determined by the critical value 1.925, the sample data significantly disagrees with the null hypothesis.



P-value Calculation

Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ 

P-value = 0.023 or 2.3% (Answers will vary)

Since the p-value was less than our significance level, it is unlikely for our sample data to have occurred because of sampling variability.

Since the p-value was less than our significance level, we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that the location of a home (state) is related to the price of a home.

ANOVA Table				
	df	SS	MS	F
Groups	3	3721834.4	1240611.5	2.747
Error	116	52382916.4	451576.9	
Total	119	56104750.8		

The variance between the groups (1240611.5) is significantly greater than the variance within the groups (451576.9). We know it is significant since our F-test statistic fell in the tail determined by the critical value.

The P-value was lower than the significance level. This indicates that the categorical variable (state) is related to the quantitative variable (price).



Section 4C Odd Answers

	Z-test stat	Sentence to explain Z-test statistic.	Critical Value	Does the Z-test statistic fall in a tail determined by a critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	-1.835	The sample proportion from group 1 was 1.835 standard errors below the sample proportion from group 2.	± 1.645	Yes. In Tail	Yes. Sig. disagree
2.	+0.974		+2.576		
3.	-1.226	The sample proportion from group 1 was 1.226 standard errors below the sample proportion from group 2.	-1.96	No. Not in tail.	No. Does not sig disagree
4.	-3.177		± 1.96		
5.	+2.244	The sample proportion from group 1 was 2.224 standard errors above the sample proportion from group 2.	+1.645	Yes. In Tail	Yes. Sig. disagree
6.	+1.448		± 2.576		
7.	-0.883	The sample proportion from group 1 was 0.883 standard errors below the sample proportion from group 2.	-2.576	No. Not in tail.	No. Does not sig disagree
8.	+1.117		+1.96		
9.	+2.139	The sample proportion from group 1 was 2.139 standard errors above the sample proportion from group 2.	± 2.576	No. Not in tail.	No. Does not sig disagree
10.	-0.199		-1.645		

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.728	72.8%	If H_0 is true, there is a 72.8% probability of getting the sample data or more extreme because of sampling variability.	10%	0.10	Could be random chance	Fail to reject H_0
12.	0.0421			1%			
13.	2.11×10^{-4}	0.0211%	If H_0 is true, there is a 0.0211% probability of getting the sample data or more extreme because of sampling variability.	5%	0.05	Unlikely to be random chance	Reject H_0
14.	0.0033			1%			
15.	0.176	17.6%	If H_0 is true, there is a 17.6% probability of getting the sample data or more extreme because of sampling variability.	5%	0.05	Could be random chance	Fail to reject H_0
16.	0			10%			



17.	0.0628	6.28%	If H_0 is true, there is a 6.28% probability of getting the sample data or more extreme because of sampling variability.	5%	0.05	Could be random chance	Fail to reject H_0
18.	0.277			10%			
19.	3.04×10^{-6}	0.000304%	If H_0 is true, there is a 0.000304% probability of getting the sample data or more extreme because of sampling variability.	1%	0.01	Unlikely to be random chance	Reject H_0
20.	0			5%			

21. A random sample is selecting people randomly from a population. It is used to help eliminate bias and make the sample data more representative of the population. Random assignment is separating a group of people in an experiment into two or more groups randomly. This makes the groups alike, controls confounding variables and helps prove cause and effect.

23.

Two-population Proportion Assumptions (Independent groups for Experiment)

- The two categorical samples should be randomly assigned from the people or objects in the experiment.
- Data values within each sample should be independent of each other.
- Data values between the samples should be independent of each other.
- Both samples should be at least ten successes and at least ten failures.

25.

To prove cause and effect, we would need to set up a controlled experiment. Randomly assign people in the experiment into two groups and make sure they meet the assumptions. The groups should be alike in order to control confounding variables. If the P-value is low and the groups are significantly different, this may indicate cause and effect.

27.

π_1 : Population proportion of marijuana users that use other drugs

π_2 : Population proportion of non-marijuana users that use other drugs

$H_0 : \pi_1 = \pi_2$ (Using marijuana is not related to using other drugs)

$H_0 : \pi_1 > \pi_2$ (Using marijuana is related to using other drugs) CLAIM

Assumptions

Random Samples? Yes both samples were collected randomly.

Individuals within and between the samples are independent? Probably yes. These were small random samples from a large population. The marijuana and non-marijuana users are unlikely to be related.

Both samples have at least 10 success and at least 10 failures? Yes. Both samples passed. Sample 1 had 87 success and $213 - 87 = 126$ failures. Sample 2 had 26 success and $219 - 26 = 193$ failures.

Z-test statistic = 6.850

The sample proportion for group 1 was 6.850 standard errors above the sample proportion for group 2.



The sample proportions are significantly different since the test statistic fell in the tail determined by the critical value.

P-value = 0.000000000036839 = 0.00000000036839% \approx 0%

If the null hypothesis is true, then there is about a 0% probability of getting the sample data or more extreme because of sampling variability.

The P-value is close to zero, so it is very unlikely that the sample data occurred because of sampling variability.

Since the P-value is less than the significance level, we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that the proportion of marijuana users that use other drugs is higher than for non-marijuana users. This also indicates that using marijuana is related to using other drugs.

29.

π_1 : Population proportion of married people that are unhappy

π_2 : Population proportion of non-married people that are unhappy

$H_0 : \pi_1 = \pi_2$ (Being married is NOT related to being unhappy)

$H_0 : \pi_1 < \pi_2$ (Being married is related to being unhappy) CLAIM

Assumptions

Random Samples? Yes both samples were collected randomly.

Individuals within and between the samples are independent? Probably yes. These were small random samples from a large population. The married, single and divorced people are unlikely to be related.

Both samples have at least 10 success and at least 10 failures? Yes. Both samples passed. Sample 1 had 74 success and $200 - 74 = 126$ failures. Sample 2 had 97 success and $200 - 97 = 103$ failures.

Z-test statistic = -2.325

The sample proportion for group 1 was 2.325 standard errors below the sample proportion for group 2.

The sample proportions are significantly different since the test statistic fell in the tail determined by the critical value.

P-value = 0.0100 = 1.0%

If the null hypothesis is true, then there is about a 1.0% probability of getting the sample data or more extreme because of sampling variability.

The P-value is less than the significance level, so it is unlikely that the sample data occurred because of sampling variability.

Since the P-value is less than the significance level, we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that married people have a lower percentage of unhappiness than non-married people. This also indicates that being married or not is related to being unhappy.

31.

π_1 : Population proportion of women with a normal BMI

π_2 : Population proportion of men with a normal BMI

$H_0 : \pi_1 = \pi_2$ (Gender is NOT related to having a normal BMI)

$H_0 : \pi_1 < \pi_2$ (Gender is related to having a normal BMI) CLAIM



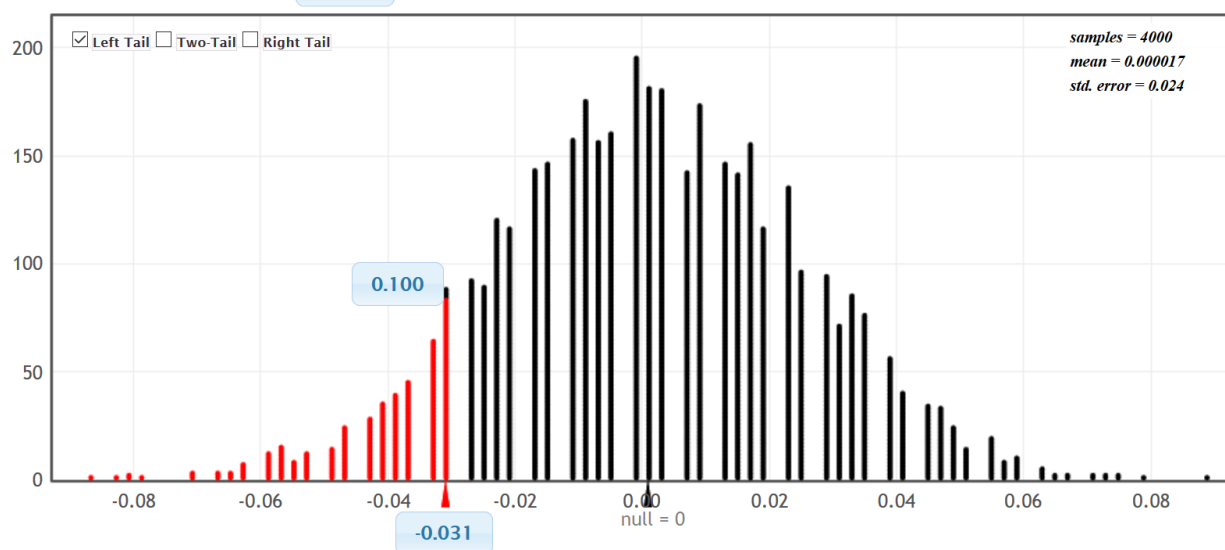
Original Sample

Group	Count	Sample Size	Proportion
Group 1	198	760	0.261
Group 2	273	745	0.366
Group 1-Group 2	-75	n/a	-0.106

The sample proportion difference is -0.106 .

Tail determined by the simulation and the significance level. (Simulations and tails will vary)

Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ Null Hypothesis: $p_1 = p_2$

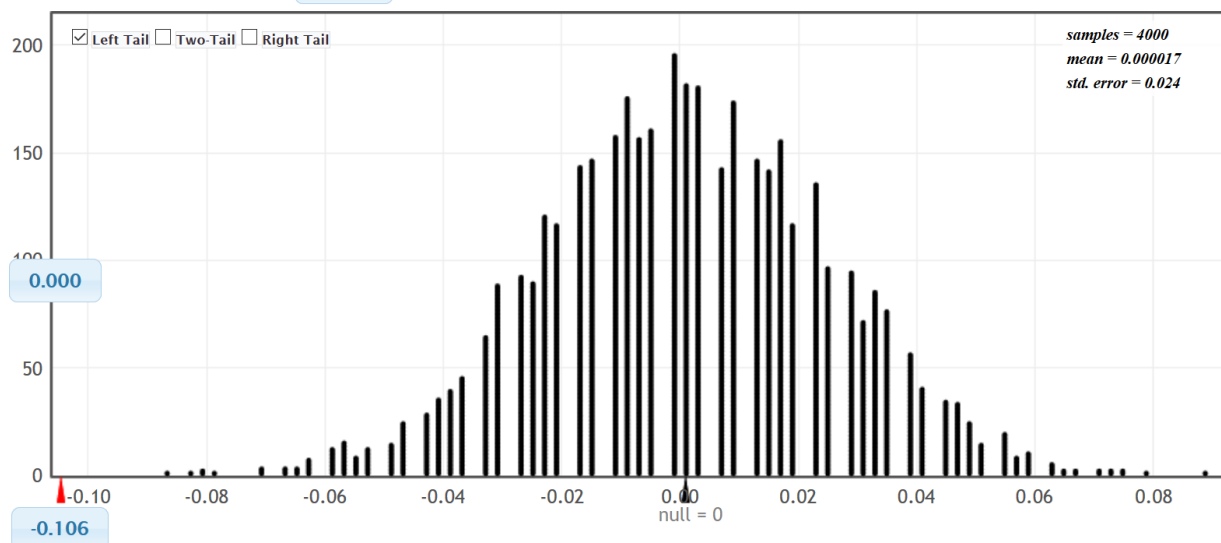


Notice that the original sample difference of -0.106 does fall in the tail determined by the significance level and the simulation. This implies that the sample proportion for women was significantly lower than the sample proportion for men.



P-value determined by the simulation and original sample difference

Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ Null Hypothesis: $p_1 = p_2$



P-value = 0 = 0% (answers will vary)

If the null hypothesis is true, there is a 0% probability of getting this sample data or more extreme because of sampling variability.

It is very unlikely that this sample data occurred because of sampling variability.

Since the P-value is less than the significance level, we will reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that percentage of women with normal BMI is lower than the percentage of men. This also implies that having a normal BMI is related to gender.

33.

π_1 : Population proportion of smoking women that are able to get pregnant

π_2 : Population proportion of non-smoking women that are able to get pregnant

$H_0 : \pi_1 = \pi_2$ (Smoking is NOT related to getting pregnant)

$H_0 : \pi_1 < \pi_2$ (Smoking is related to getting pregnant) CLAIM

Original Sample

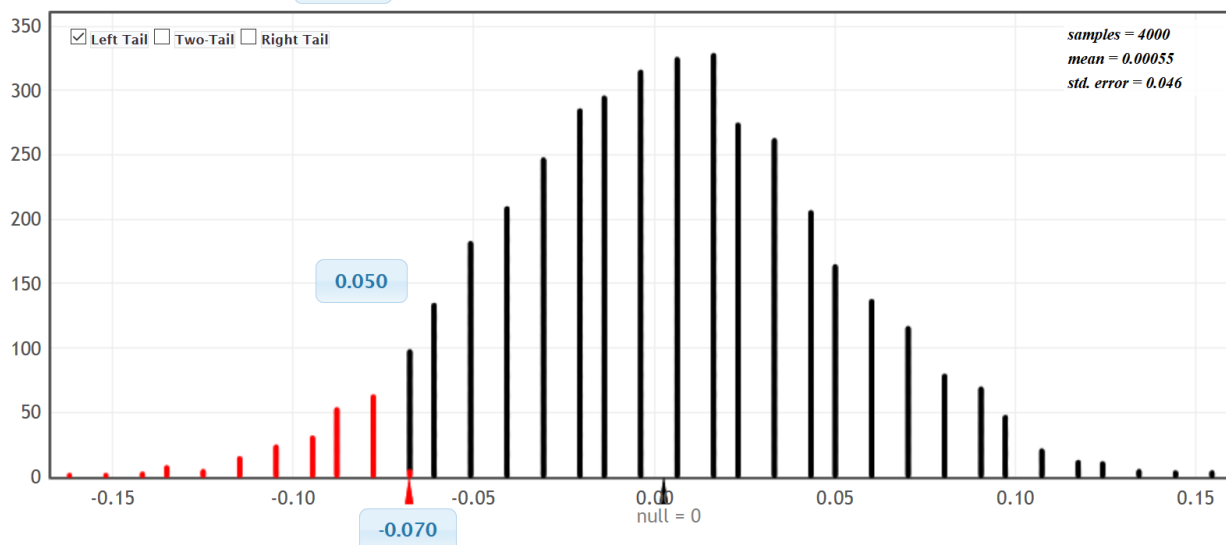
Group	Count	Sample Size	Proportion
Group 1	38	135	0.281
Group 2	206	543	0.379
Group 1-Group 2	-168	n/a	-0.098

The sample proportion difference is -0.098.



Tail determined by the simulation and the significance level. (Simulations and tails will vary)

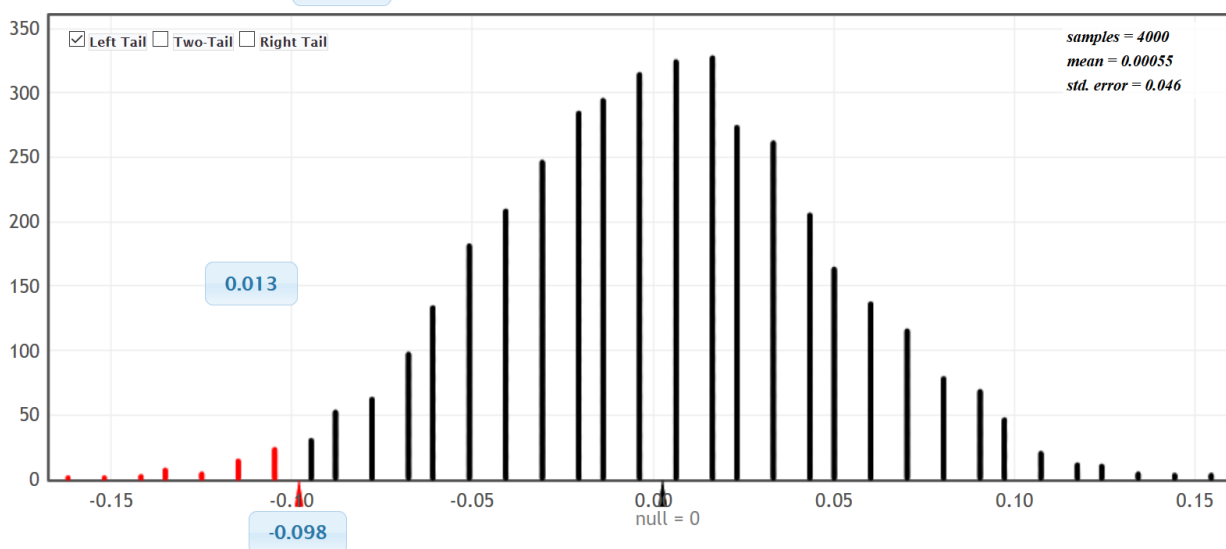
Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ Null Hypothesis: $p_1 = p_2$



Notice that the original sample difference of -0.098 does fall in the tail determined by the significance level and the simulation. This implies that the sample proportion for smokers was significantly lower than the sample proportion for non-smokers.

P-value determined by the simulation and original sample difference

Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ Null Hypothesis: $p_1 = p_2$



P-value = 0.013 = 1.3% (answers will vary)

If the null hypothesis is true, there is a 1.3% probability of getting this sample data or more extreme because of sampling variability.

It is very unlikely that this sample data occurred because of sampling variability.

Since the P-value is less than the significance level, we will reject the null hypothesis.



Conclusion: There is significant evidence to support the claim that percentage of smokers that are able to get pregnant is lower than the percentage of non-smokers. This also implies that getting pregnant is related to smoking or not.

Section 4D Odd Answers

	χ^2 -test stat	Sentence to explain χ^2 -test statistic.	Critical Value	Does the χ^2 -test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	+28.573	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 28.573.	+9.117	Yes. In tail.	Yes. Significantly disagrees.
2.	+1.226		+7.113		
3.	+2.137	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 2.137	+5.521	No. Not in tail.	No. Does not significantly disagree.
4.	+14.415		+6.114		
5.	+3.718	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 3.718	+7.182	No. Not in tail.	No. Does not significantly disagree.
6.	+0.891		+3.994		
7.	+51.652	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 51.652	+14.881	Yes. In tail.	Yes. Significantly disagrees.
8.	+1.185		+4.181		
9.	+2.442	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 2.442	+8.619	No. Not in tail.	No. Does not significantly disagree.
10.	+14.133		+10.336		

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.0006	0.06%	If H_0 is true, there is a 0.06% probability of getting the sample data or more extreme by sampling variability.	10%	0.1	Unlikely	Reject H_0



12.	0.042			1%			
13.	9.16×10^{-7}	0.0000916%	If H_0 is true, there is a 0.0000916% probability of getting the sample data or more extreme by sampling variability.	5%	0.05	Unlikely	Reject H_0
14.	0.739			1%			
15.	0.0035	0.35%	If H_0 is true, there is a 0.35% probability of getting the sample data or more extreme by sampling variability.	5%	0.05	Unlikely	Reject H_0
16.	0			10%			
17.	0.419	41.9%	If H_0 is true, there is a 41.9% probability of getting the sample data or more extreme by sampling variability.	5%	0.05	Could be	Fail to reject H_0
18.	0.0274			10%			
19.	3.77×10^{-5}	0.00377%	If H_0 is true, there is a 0.00377% probability of getting the sample data or more extreme by sampling variability.	1%	0.01	Unlikely	Reject H_0
20.	0.067			5%			

21.

Goodness of Fit Degrees of Freedom = $K - 1$

(K is the number of groups)

23.

For each group, the computer subtracts the expected count (H_0) from the observed count (Sample). It then squares the differences to get rid of negatives. It then divides each square by the expected count. Lastly, it adds up this calculation for each group to get the total chi-squared.

25.

If the observed sample counts and the expected counts from the H_0 were close, the differences would be very small. That would cause the overall chi-squared to be small.



27.

$$H_0 : \pi_{\text{bike}} = 0.01, \pi_{\text{carpool}} = 0.1, \pi_{\text{drive alone}} = 0.8, \\ \pi_{\text{dropped off}} = 0.05, \pi_{\text{public transp}} = 0.02, \pi_{\text{walk}} = 0.02$$

 H_A : At least one proportion is \neq (CLAIM)

Degrees of freedom = 6 – 1 = 5

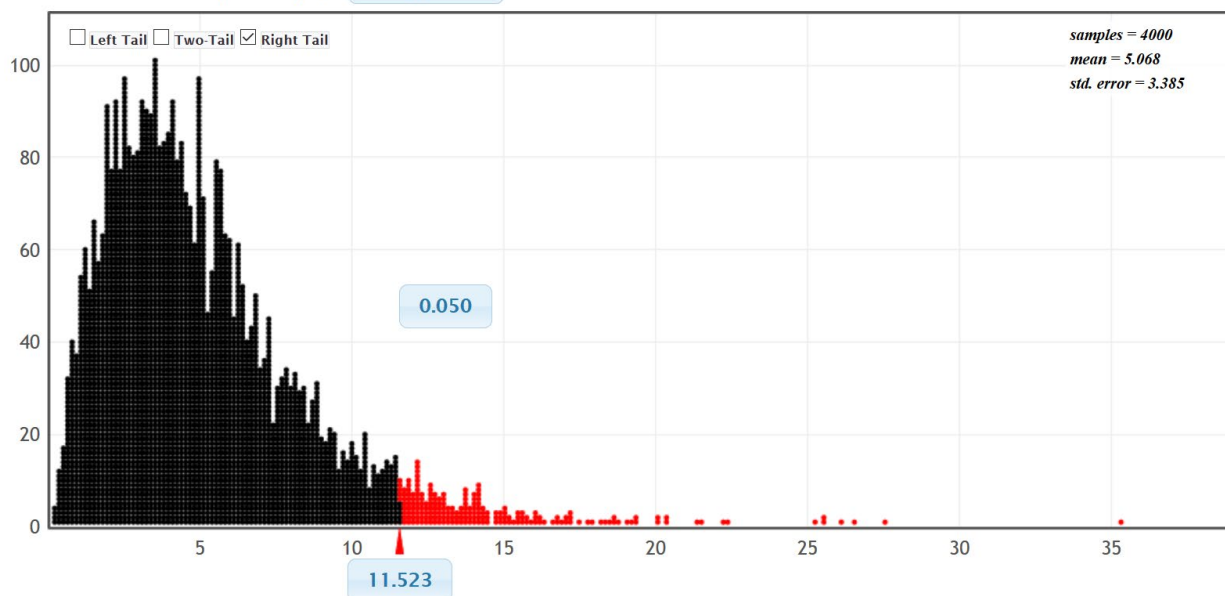
Original Sample[Show Details](#) $n = 332, \chi^2 = 3.816$

	Count
Bicycle	1
Carpool	30
Drive Alone	267
Dropped Off	18
Public Transportation	6
Walk	10

Chi-squared test stat = 3.816

The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis is 3.816.

Critical value and Tail calculation from simulation and significance level (answers will vary)

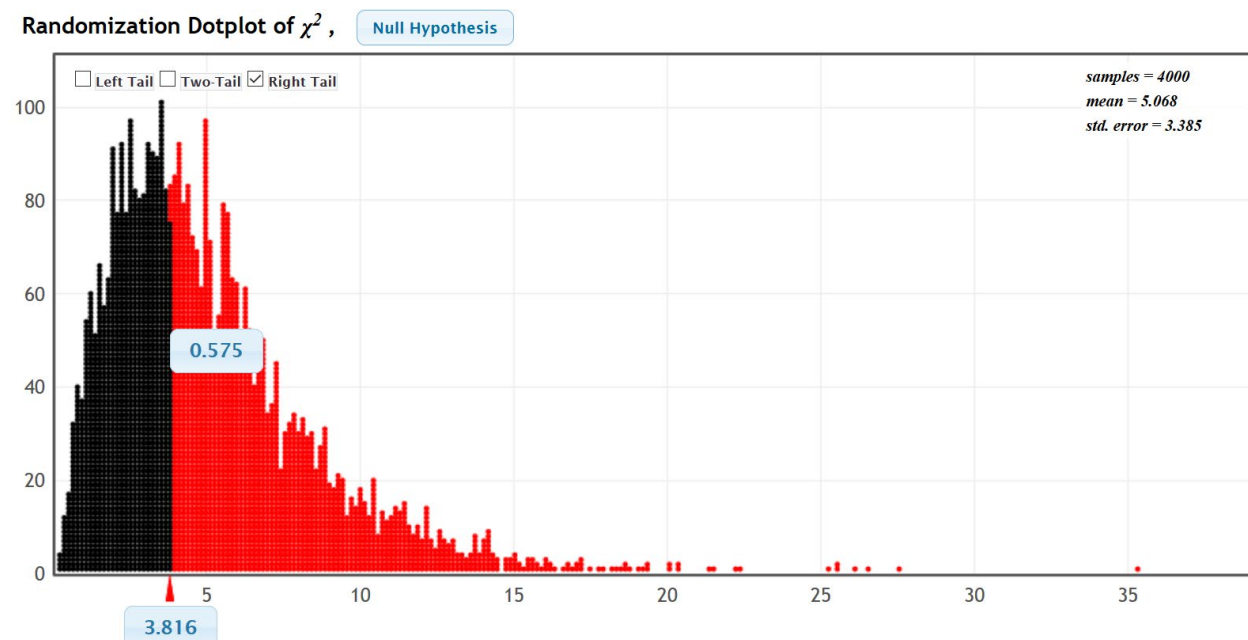
Randomization Dotplot of χ^2 ,[Null Hypothesis](#)

Critical Value = 11.523 (answers will vary)

The test statistic of 3.816 does not fall in the tail determined by the critical value. So the sample data does not significantly disagree with the null hypothesis. Also the observed sample counts are not significantly different than the expected counts from the null hypothesis.



P-value calculation from the test statistic and simulation (answers will vary)



P-value = 0.575 = 57.5% (answers will vary)

If the null hypothesis was true, then there is a 57.5% probability of getting the sample data or more extreme by sampling variability.

The P-value is much higher than the 5% significance level. So this sample data could have occurred simply by sampling variability.

Fail to reject the null hypothesis.

Conclusion: There is not significant evidence to support the claim that the percentages listed by the employee are wrong. They might be correct, but we do not have evidence.

This problem is the second type of goodness of fit test and is not designed to explore relationships. The P-value in this case was trying to test the accuracy of the given percentages, not tell if the variables are related. The percentages being so vastly different does indicate that type of transportation is probably related to the percentage of people that use that type.

29.

$$H_0 : \pi_{\text{caucasian}} = 0.54, \pi_{\text{african american}} = 0.18, \pi_{\text{hispanic american}} = 0.12, \\ \pi_{\text{asian american}} = 0.15, \pi_{\text{other}} = 0.01$$

$$H_A : \text{At least one proportion is } \neq \text{ (CLAIM)}$$

$$\text{Degrees of freedom} = 5 - 1 = 4$$



Original Sample

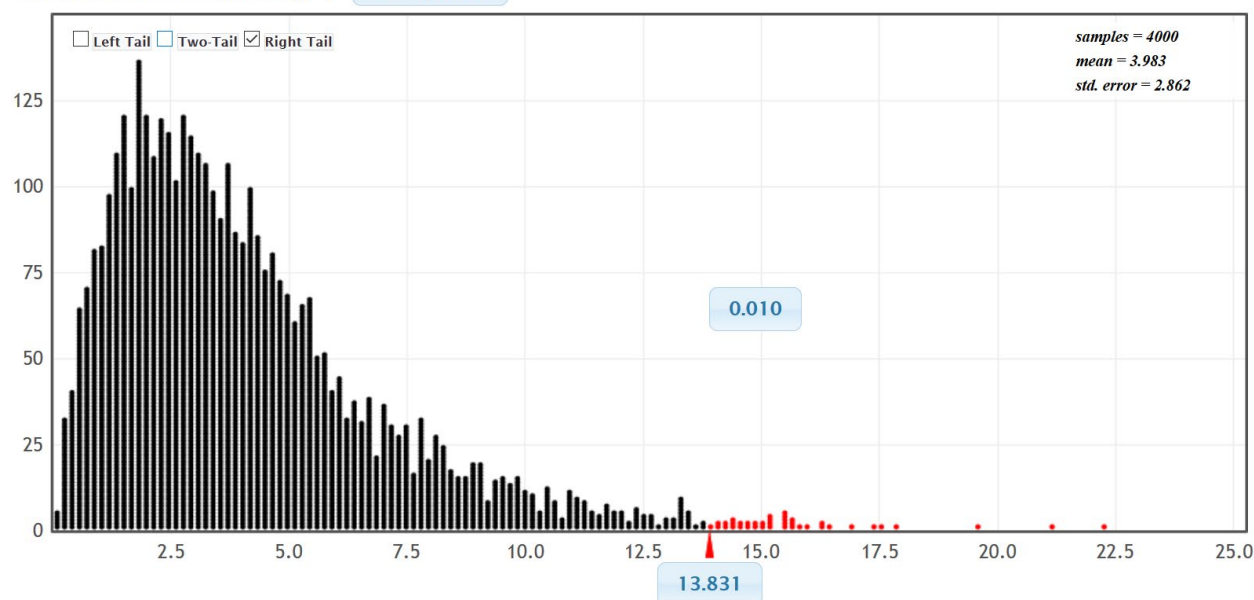
[Show Details](#) $n = 1453, \chi^2 = 357.362$

	Count
Caucasian	780
African American	117
Hispanic American	114
Asian American	384
Other	58

Chi-squared test stat = 357.362

The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis is 357.362.

Critical value and Tail calculation from simulation and significance level (answers will vary)

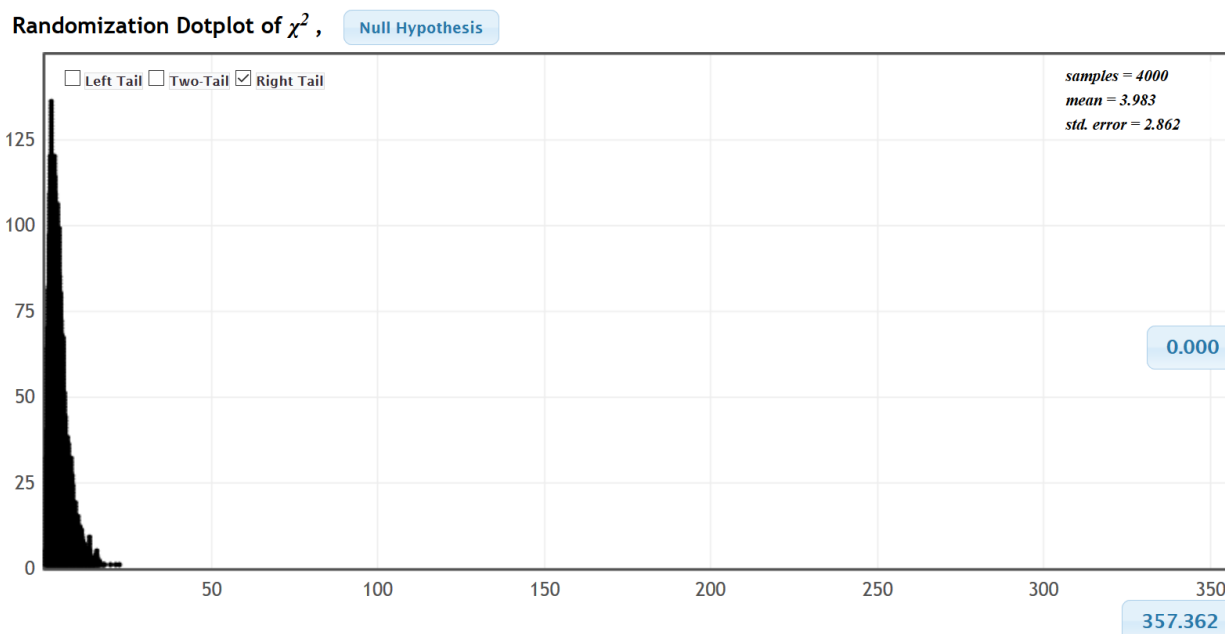
Randomization Dotplot of χ^2 ,[Null Hypothesis](#)

Critical Value = 13.831 (answers will vary)

The test statistic of 357.362 is way out in the far tail determined by the critical value. So the sample data significantly disagrees with the null hypothesis. Also the observed sample counts are significantly different than the expected counts from the null hypothesis.



P-value calculation from the test statistic and simulation (answers will vary)



P-value = 0 = 0% (answers may vary)

If the null hypothesis was true, then there is a 0% probability of getting the sample data or more extreme by sampling variability.

The P-value is much lower than the 1% significance level. It is highly unlikely for this sample data to have occurred simply by sampling variability.

Reject the null hypothesis.

Conclusion: There is significant evidence to support the claim that the juries from this county are not representing the demographic.

This problem is the second type of goodness of fit test and is not designed to explore relationships. The P-value in this case was trying to test the accuracy of the given percentages, not tell if the variables are related. The percentages being so vastly different does indicate that race is most likely related to the percentages.

31.

$$H_0 : \pi_{\text{monday}} = \pi_{\text{tuesday}} = \pi_{\text{wednesday}} = \pi_{\text{thursday}} = \pi_{\text{friday}} = \pi_{\text{saturday}} = \pi_{\text{sunday}}$$

H_A : At least one proportion is \neq (CLAIM)

$$\text{Degrees of freedom} = 7 - 1 = 6$$

$$\text{Chi-squared test stat} = 7.5478$$

The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis is 7.5478.

$$\text{Critical Value} = 16.8119$$

The test statistic of 7.5478 does not fall in the tail determined by the critical value. So the sample data does NOT significantly disagree with the null hypothesis. Also the observed sample counts are NOT significantly different than the expected counts from the null hypothesis.

P-value = 0.2731 = 27.31% (answers may vary)



If the null hypothesis was true, then there is a 27.31% probability of getting the sample data or more extreme by sampling variability.

The P-value is higher than the 1% significance level. This sample data could have occurred simply by sampling variability.

Fail to reject the null hypothesis.

Conclusion: There is not significant evidence to support the claim that the probability of having a fatal car accident is different on the various days of the week. This also implies that the day of the week is probably not related to having a fatal car accident.

Section 4E Odd Answers

1. If conditional proportions are significantly different from one group to another, it indicates that the categorical variable that decides the groups is probably related to the categorical variable that the proportions came from.
2. If conditional proportions are almost the same from one group to another, it indicates that the categorical variable that decides the groups is probably not related to the categorical variable that the proportions came from.

#3-8

Counts Table [Switch Variables](#)

Do you smoke cigarettes? \ What type of transportation do you take to campus?	Drive alone	Dropped off by someone	Carpool	Bicycle	Public transportation	Walk	Total
No	236	17	28	1	6	10	298
Yes	26	1	2	0	0	0	29
Total	262	18	30	1	6	10	327

Proportions [Row](#) [Column](#) [Overall](#)

Do you smoke cigarettes? \ What type of transportation do you take to campus?	Drive alone	Dropped off by someone	Carpool	Bicycle	Public transportation	Walk	Total
No	0.722	0.052	0.086	0.0031	0.018	0.031	0.911
Yes	0.08	0.0031	0.0061	0	0	0	0.089
Total	0.801	0.055	0.092	0.0031	0.018	0.031	1

3. At the end of yes smoking row in the overall chart we see the answer is 0.089 or 8.9%. We could also have calculated this by dividing the amount of smokers (29) by the grand total (327).

5. In the overall chart where no smoking and carpool meet, we see the answer is 0.086 or 8.6%. We could also have calculated this by dividing the cell where no smoking and carpool meet (28) by the grand total (327).

7. We will use the union "or" formula. These individual proportions we can get from the overall chart.

$$P(\text{no smoke OR dropped off}) = P(\text{no smoke}) + P(\text{dropped off}) - P(\text{no smoke and dropped off}) = 0.911 + 0.055 - 0.052 = 0.914 \text{ or } 91.4\%$$



#9-10

Counts Table [Switch Variables](#)

Do you smoke cigarettes? \ What type of transportation do you take to campus?	Drive alone	Dropped off by someone	Carpool	Bicycle	Public transportation	Walk	Total
No	236	17	28	1	6	10	298
Yes	26	1	2	0	0	0	29
Total	262	18	30	1	6	10	327

Proportions [Row](#) [Column](#) [Overall](#)

Do you smoke cigarettes? \ What type of transportation do you take to campus?	Drive alone	Dropped off by someone	Carpool	Bicycle	Public transportation	Walk	Total
No	0.901	0.944	0.933	1	1	1	0.911
Yes	0.099	0.056	0.067	0	0	0	0.089
Total	1	1	1	1	1	1	1

Proportions [Row](#) [Column](#) [Overall](#)

Do you smoke cigarettes? \ What type of transportation do you take to campus?	Drive alone	Dropped off by someone	Carpool	Bicycle	Public transportation	Walk	Total
No	0.792	0.057	0.094	0.0034	0.02	0.034	1
Yes	0.897	0.034	0.069	0	0	0	1
Total	0.801	0.055	0.092	0.0031	0.018	0.031	1

9. Since the conditions are smoking and not smoking and they are a row, we will use the row chart for conditional %.

$P(\text{carpool given smoke}) = 0.069 = 6.9\%$. We can also calculate this by dividing the # of carpool smokers (2) by the total number of smokers (29).

$P(\text{carpool given not smoke}) = 0.094 = 9.4\%$. We can also calculate this by dividing the # of carpool non-smokers (28) by the total number of non-smokers (298).

These appear significantly different. (36% increase) So smoking or not does appear to be related to carpooling.

#11-16

Rows: text and drive or not

Columns: car accident or not



Counts Table Switch Variables

Do you text while you drive? \ Have you been involved a car accident (as the driver) within the past two years?	No	Yes	Total
Yes	89	43	132
No	159	40	199
Total	248	83	331

Proportions Row Column Overall

Do you text while you drive? \ Have you been involved a car accident (as the driver) within the past two years?	No	Yes	Total
Yes	0.269	0.13	0.399
No	0.48	0.121	0.601
Total	0.749	0.251	1

11. At the end of yes texting and driving row in the overall chart we see the answer is 0.399 or 39.9%. We can also calculate this by dividing the total number of texting while driving (132) by the grand total (331).

13. In the overall chart where yes text and drive and yes car accident meet, we see the answer is 0.130 or 13.0%. We can also calculate this by dividing the total number of texting while driving car accidents (43) by the grand total (331).

15. We will use the union “or” formula. These individual proportions we can get from the overall chart.

$P(\text{yes text drive OR no car accident}) = P(\text{yes text drive}) + P(\text{no car accident}) - P(\text{yes text drive AND no car accident}) = 0.399 + 0.749 - 0.269 = 0.879$ or 87.9%

#17-18

Counts Table Switch Variables

Do you text while you drive? \ Have you been involved a car accident (as the driver) within the past two years?	No	Yes	Total
Yes	89	43	132
No	159	40	199
Total	248	83	331

Proportions Row Column Overall

Do you text while you drive? \ Have you been involved a car accident (as the driver) within the past two years?	No	Yes	Total
Yes	0.674	0.326	1
No	0.799	0.201	1
Total	0.749	0.251	1



Proportions

Row

Column

Overall

Do you text while you drive? \ Have you been involved a car accident (as the driver) within the past two years?	No	Yes	Total
Yes	0.359	0.518	0.399
No	0.641	0.482	0.601
Total	1	1	1

17. Since the conditions are text & drive and not text & drive are rows, we will use the row chart.

$P(\text{car accident given text\&drive}) = 0.326 = 32.6\%$. We can also calculate this by dividing the number of text and drive car accidents (43) by the total number of texting and driving (132).

$P(\text{car accident given not text\&drive}) = 0.201 = 20.1\%$. We can also calculate this by dividing the number of no text and drive car accidents (40) by the total number of not texting and driving (199).

These appear significantly different. (62.2% increase) So being in a car accident does appear to be related to texting and driving or not.

#19-24

Counts Table

Switch Variables

Do you have at least one tattoo? \ Which social media do you use the most?	Snapchat	Other	Facebook	Instagram	Twitter	Total
No	60	21	56	83	21	241
Yes	11	6	19	41	8	85
Total	71	27	75	124	29	326

Proportions

Row

Column

Overall

Do you have at least one tattoo? \ Which social media do you use the most?	Snapchat	Other	Facebook	Instagram	Twitter	Total
No	0.184	0.064	0.172	0.255	0.064	0.739
Yes	0.034	0.018	0.058	0.126	0.025	0.261
Total	0.218	0.083	0.23	0.38	0.089	1

19. At the end of yes tattoo row in the overall chart we see the answer is 0.261 or 26.1%. We can also calculate this by dividing the total number of students with tattoos (85) by the grand total (326).

21. In the overall chart where Facebook and no tattoo meet, we see the answer is 0.172 or 17.2%. We can also calculate this by dividing the number of Facebook no tattoo students (56) by the grand total (326).

23. We will use the union "or" formula. These individual proportions we can get from the overall chart.

$$P(\text{yes tattoo OR Instagram}) = P(\text{yes tattoo}) + P(\text{Instagram}) - P(\text{yes tattoo AND Instagram}) = \\ = 0.261 + 0.38 - 0.126 = 0.515 \text{ or } 51.5\%$$



#25-26

Counts Table[Switch Variables](#)

Do you have at least one tattoo? \ Which social media do you use the most?	Snapchat	Other	Facebook	Instagram	Twitter	Total
No	60	21	56	83	21	241
Yes	11	6	19	41	8	85
Total	71	27	75	124	29	326

Proportions[Row](#)[Column](#)[Overall](#)

Do you have at least one tattoo? \ Which social media do you use the most?	Snapchat	Other	Facebook	Instagram	Twitter	Total
No	0.249	0.087	0.232	0.344	0.087	1
Yes	0.129	0.071	0.224	0.482	0.094	1
Total	0.218	0.083	0.23	0.38	0.089	1

Proportions[Row](#)[Column](#)[Overall](#)

Do you have at least one tattoo? \ Which social media do you use the most?	Snapchat	Other	Facebook	Instagram	Twitter	Total
No	0.845	0.778	0.747	0.669	0.724	0.739
Yes	0.155	0.222	0.253	0.331	0.276	0.261
Total	1	1	1	1	1	1

25. Since the conditions of tattoo and no tattoo are in a row, we will use the row chart for conditional %.

$P(\text{Twitter given tattoo}) = 0.094 = 9.4\%$. We can also calculate this by dividing the number of Twitter with tattoo students (8) by the total number of students with tattoo (85).

$P(\text{Twitter given no tattoo}) = 0.087 = 8.7\%$. We can also calculate this by dividing the number of Twitter with no tattoo students (21) by the total number of students with no tattoo (241).

These appear to be relatively close. (Only 8% increase) So having a tattoo or not does not appear to be related to liking Twitter.



#27-32

Counts Table Switch Variables

Cylinders \ Country	U.S.	Japan	Germany	Sweden	France	Italy	Total
8	8	0	0	0	0	0	8
4	7	6	4	1	0	1	19
5	0	0	1	0	0	0	1
6	7	1	0	1	1	0	10
Total	22	7	5	2	1	1	38

Proportions Row Column Overall

Cylinders \ Country	U.S.	Japan	Germany	Sweden	France	Italy	Total
8	0.211	0	0	0	0	0	0.211
4	0.184	0.158	0.105	0.026	0	0.026	0.5
5	0	0	0.026	0	0	0	0.026
6	0.184	0.026	0	0.026	0.026	0	0.263
Total	0.579	0.184	0.132	0.053	0.026	0.026	1

27. At the end of the Germany column in the overall chart we see the answer is 0.132 or 13.2%. We could also calculate this by dividing the total number of cars from Germany (5) by the grand total (38).

29. In the overall chart where Japan and four cylinders meet, we see the answer is 0.158 or 15.8%. We could also have calculated this by dividing the number of four cylinder cars from Japan (6) by the grand total (38).

31. We will use the union “or” formula. These individual proportions we can get from the overall chart.

$$P(\text{Germany OR Six Cylinders}) = P(\text{Germany}) + P(\text{Six Cylinders}) - P(\text{Germany AND Six Cylinders}) = \\ = 0.132 + 0.263 - 0 = 0.395 \text{ or } 39.5\%$$

#33-34

Counts Table Switch Variables

Cylinders \ Country	U.S.	Japan	Germany	Sweden	France	Italy	Total
8	8	0	0	0	0	0	8
4	7	6	4	1	0	1	19
5	0	0	1	0	0	0	1
6	7	1	0	1	1	0	10
Total	22	7	5	2	1	1	38



Proportions

Row

Column

Overall

Cylinders \ Country	U.S.	Japan	Germany	Sweden	France	Italy	Total
8	1	0	0	0	0	0	1
4	0.368	0.316	0.211	0.053	0	0.053	1
5	0	0	1	0	0	0	1
6	0.7	0.1	0	0.1	0.1	0	1
Total	0.579	0.184	0.132	0.053	0.026	0.026	1

Proportions

Row

Column

Overall

Cylinders \ Country	U.S.	Japan	Germany	Sweden	France	Italy	Total
8	0.364	0	0	0	0	0	0.211
4	0.318	0.857	0.8	0.5	0	1	0.5
5	0	0	0.2	0	0	0	0.026
6	0.318	0.143	0	0.5	1	0	0.263
Total	1	1	1	1	1	1	1

33. Since the conditions of Japan and Germany are in columns, we will use the column chart for conditional %.

$P(\text{Four Cylinders given Japan}) = 0.857 = 85.7\%$. We can also calculate this by dividing the number of four cylinder cars from Japan (6) by the total number of cars from Japan (7).

$P(\text{Four Cylinders given Germany}) = 0.8 = 80\%$. We can also calculate this by dividing the number of four cylinder cars from Germany (4) by the total number of cars from Germany (5).

These appear to be relatively close. (Only 7.1% increase)

So the country (Japan and Germany) is probably not related to having four cylinders.

Section 4F Odd Answers

	χ^2 -test stat	Sentence to explain χ^2 -test statistic.	Critical Value	Does the χ^2 -test statistic fall in a tail determined by the critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	+1.573	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 1.573.	+4.117	No. Not in tail.	No. Does not significantly disagree.
2.	+6.226		+5.118		
3.	+2.144	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 2.144.	+4.121	No. Not in tail.	No. Does not significantly disagree.



4.	+3.415		+5.091		
5.	+13.718	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 13.718.	+7.189	Yes. In tail.	Yes. Significantly disagrees.
6.	+0.972		+4.812		
7.	+31.652	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 31.652.	+12.557	Yes. In tail.	Yes. Significantly disagrees.
8.	+11.185		+5.181		
9.	+25.443	The sum of the averages of the squares of the differences between the observed sample values and the expected values from the null hypothesis is 25.443.	+7.008	Yes. In tail.	Yes. Significantly disagrees.
10.	+1.133		+8.336		

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.263	26.3%	If H_0 is true, there is a 26.3% probability of getting the sample data or more extreme by sampling variability	10%	0.1	Could be	Fail to reject H_0
12.	0.0042			1%			
13.	5.22×10^{-4}	0.0522%	If H_0 is true, there is a 0.0522% probability of getting the sample data or more extreme by sampling variability	5%	0.05	Unlikely	Reject H_0
14.	0.0639			1%			
15.	0	0%	If H_0 is true, there is a 0% probability of getting the sample data or more extreme by sampling variability	5%	0.05	Unlikely	Reject H_0
16.	0.539			10%			
17.	0.0419	4.19%	If H_0 is true, there is a 4.19% probability of getting the sample data or more extreme by sampling variability	5%	0.05	Unlikely	Reject H_0
18.	0.0027			10%			



19.	7.73×10^{-8}	0.00000773%	If H_0 is true, there is a 0.00000773% probability of getting the sample data or more extreme by sampling variability	1%	0.01	Unlikely	Reject H_0
20.	0.674			5%			

21.

To perform a categorical association test in Statcato, follow the following steps.

If you have a contingency table, then type the contingency table in data sheet. Column titles will be in the gray where it says VAR. Click on the “statistics” menu and then click “multinomial experiments”. Then click on “chi-square contingency table”. Click on the columns that contain your counts in your contingency table and press “add to list”. Then put in the significance level and press “OK”.

If you have two raw categorical data sets, copy and paste them into the data sheet. Titles should be in the gray where it says VAR. Click on the “statistics” menu and then click “Cross Tabulation and Chi-square”. Click on the row and column, the significance level and then press “OK”.

23.

Categorical Association Test Assumptions (one random sample)

- The categorical sample or samples should be collected randomly or be representative of the population.
- Data values within each sample should be independent of each other.
- The expected counts from the null hypothesis should be at least five.

25.

$$\text{Expected Counts} = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$$

27.

If the expected counts from the null hypothesis are close to the observed sample counts, then the differences between them will be close to zero. This will make the chi-squared test statistic very small and not fall in the tail. This would indicate that the sample data (observed counts) does not significantly disagree with the null hypothesis (expected counts).

29.

H_0 : Blood type is not related to the Rh

H_A : Blood type is related to the Rh (CLAIM)

Assumptions

Random? Yes. Given

Individuals Independent? Probably. Since this was a small random sample from a very large population.

Expected Counts at least 5? No. (The expected counts were 36.03, 23.0, 16.1, 85.87, 10.97, 7.0, 4.9, and 26.13)

Notice one of the expected counts (4.9) was below 5.

χ^2 Test Statistic = 8.5522

The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis is 8.5522.



The sample data significantly disagrees with the null hypothesis since the test statistic falls in the right tail determined by the critical value 6.2514. This also indicates that the observed sample counts were significantly different than the expected counts.

$$P\text{-value} = 0.0359 = 3.59\%$$

If the null hypothesis is true, there is a 3.59% probability of getting the sample data or more extreme by sampling variability.

The P-value is lower than the 10% significance level.

It is unlikely for this data to have occurred by sampling variability.

Reject H_0 .

If the sample data had met the assumptions, then the conclusion would be: There is significant evidence to support the claim that a persons' blood type is related to the Rh. However, this data did not meet all of the assumptions, so evidence is in question.

The P-value is low, indicating that blood type and Rh are likely to be related. However, our data did not meet all of the assumptions, so our evidence for this is in question.

31.

H_0 : Health is not related to education

H_A : Health is related to education (CLAIM)

Assumptions

Random? Yes. Given

Individuals Independent? Probably. Since this was a small random sample from a very large population.

Expected Counts at least 5? Yes. (The expected counts were 148.64, 249.76, 106.91, 29.70, 502.31, 844.04, 361.29, 100.36, 77.24, 129.78, 55.55, 15.43, 161.42, 271.23, 116.10, 32.25, 86.40, 145.19, 62.15, and 17.26) All were greater than 5.

$$\chi^2 \text{ Test Statistic} = 285.0610$$

The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis is 285.0610.

The sample data significantly disagrees with the null hypothesis since the test statistic falls in the right tail determined by the critical value 21.0261. This also indicates that the observed sample counts were significantly different than the expected counts.

$$P\text{-value} = 0 = 0\%$$

If the null hypothesis is true, there is a 0% probability of getting the sample data or more extreme by sampling variability.

The P-value is lower than the 5% significance level.

It is unlikely for this data to have occurred by sampling variability.

Reject H_0 .

Conclusion: There is significant evidence to support the claim that a persons' health is related to their education.

The P-value is low, indicating that health and education are likely to be related.



33.

Detailed Sample Table					
	8	4	5	6	Total
U.S.	8 4.6 2.45	7 1.1 1.455	0 0.6 0.579	7 5.8 0.253	22
Japan	0 1.5 1.474	6 3.5 1.786	0 0.2 0.184	1 1.8 0.385	7
Germany	0 1.1 1.053	4 2.5 0.9	1 0.1 5.732	0 1.3 1.316	5
Sweden	0 0.4 0.421	1 1 0	0 0.1 0.053	1 0.5 0.426	2
France	0 0.2 0.211	0 0.5 0.5	0 0 0.026	1 0.3 2.063	1
Italy	0 0.2 0.211	1 0.5 0.5	0 0 0.026	0 0.3 0.263	1
Total	8	19	1	10	38

Observed, Expected, Contribution to χ^2

Original Sample[Show Details](#)

$$n = 38, \chi^2 = 22.267$$

	8	4	5	6	Total
U.S.	8	7	0	7	22
Japan	0	6	0	1	7
Germany	0	4	1	0	5
Sweden	0	1	0	1	2
France	0	0	0	1	1
Italy	0	1	0	0	1
Total	8	19	1	10	38

Randomization Sample[Show Details](#)

H_0 : The country a car is made in is not related to the number of cylinders.

H_A : The country a car is made in is related to the number of cylinders. (CLAIM)

Assumptions

Random? Yes. The car data was collected randomly.

Individuals Independent? Probably not. The population of types of cars is not that large and many types of cars are owned by the same company. The company may have similar numbers of cylinders in their cars.

Expected Counts at least 5? No. Most of the expected counts were below 5. This means this data would not be suitable for using the traditional chi-squared distribution. Notice the simulation does not look like it fits the chi-squared distribution very well. However this data may be used for simulation if the independence assumption and random had passed.

$$\text{Degrees of freedom} = (r - 1)(c - 1) = (6 - 1)(4 - 1) = (5)(3) = 15$$

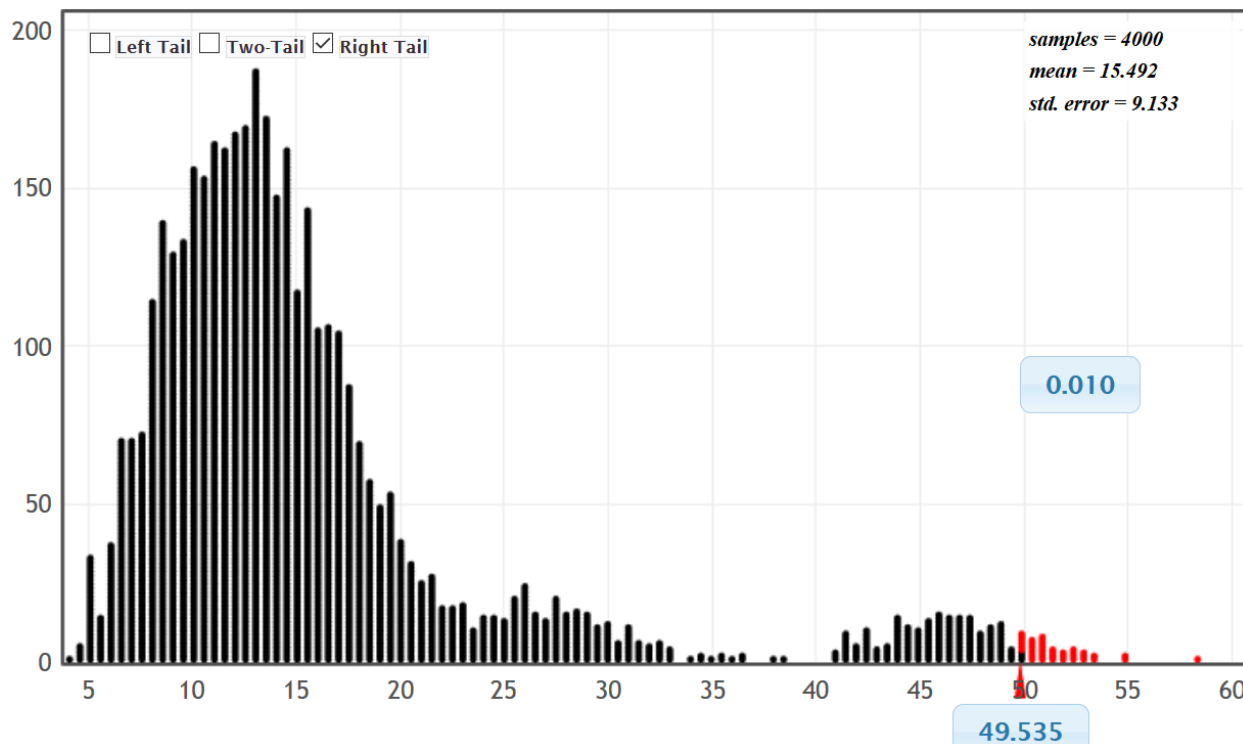
$$\chi^2 \text{ Test Statistic} = 22.267$$

The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis is 22.267.



Critical Value and tail Calculation determined by significance level (simulations will vary)

Randomization Dotplot of χ^2 , Null hypothesis: No Association



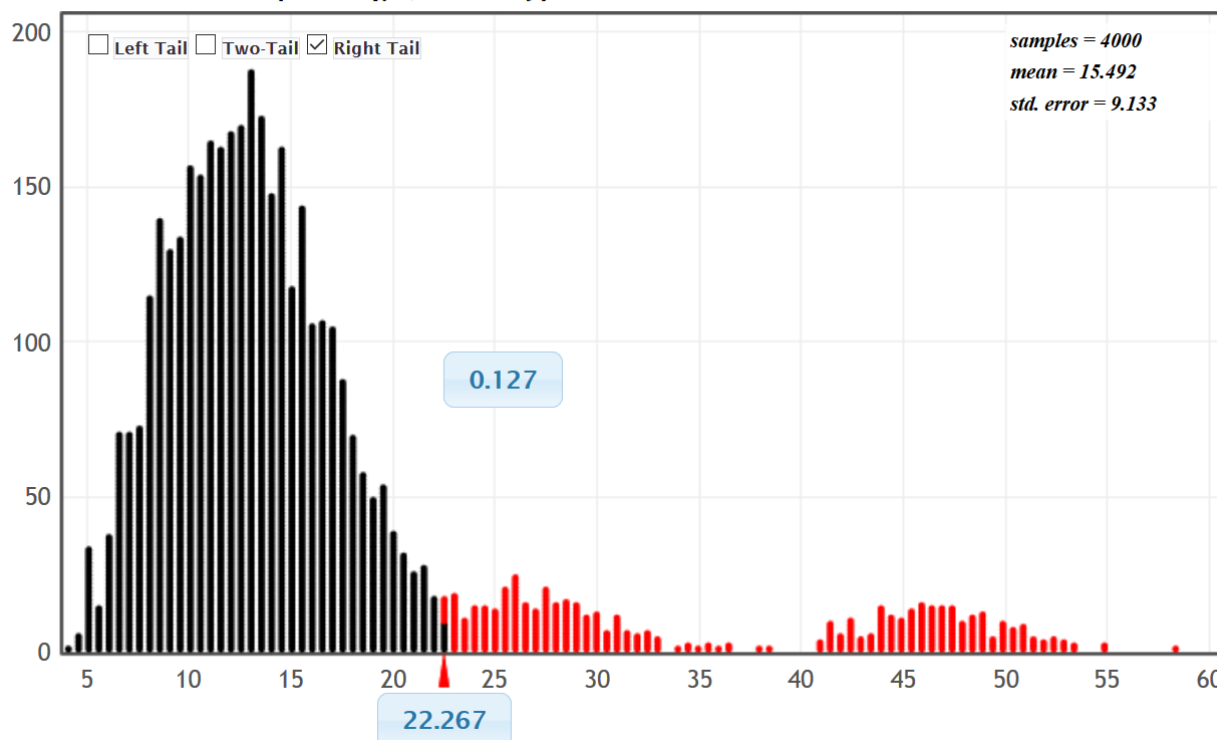
Approximate Critical value = 49.535 (answers may vary)

Notice that the Chi-square test statistic does not fall in the tail determined by the critical value. So the sample data does not significantly disagree with the null hypothesis. This also indicates that the observed sample counts were not significantly different than the expected counts.



P-value calculation with test statistic (P-values will vary)

Randomization Dotplot of χ^2 , Null hypothesis: No Association



Approximate P-value = 0.127 = 12.7% (answers will vary)

If the null hypothesis is true, there is a 12.7% probability of getting the sample data or more extreme by sampling variability.

The P-value is higher than the 1% significance level.

This sample data could have occurred by sampling variability.

Fail to reject H_0 .

Conclusion: There is NOT significant evidence to support the claim that the country a car is made in is related to the number of cylinders.

The P-value is high, indicating that the country and cylinders are likely to be not related. However, a high P-value is not evidence and this data did not pass all of the assumptions for randomized simulation.



35.

Detailed Sample Table			
	No	Yes	Total
Snapchat	60 52.5 1.075	11 18.5 3.048	71
Other	21 20 0.054	6 7 0.154	27
Facebook	56 55.4 0.0056	19 19.6 0.016	75
Instagram	83 91.7 0.82	41 32.3 2.324	124
Twitter	21 21.4 0.009	8 7.6 0.025	29
Total	241	85	326

Observed, Expected, Contribution to χ^2

Original Sample

[Show Details](#)

$$n = 326, \chi^2 = 7.531$$

	No	Yes	Total
Snapchat	60	11	71
Other	21	6	27
Facebook	56	19	75
Instagram	83	41	124
Twitter	21	8	29
Total	241	85	326

Randomization Sample

[Show Details](#)

H_0 : Tattoos are not related to social media. (CLAIM)

H_A : Tattoos are related to social media.

Assumptions

Random or representative? Yes. This data was a census of all math 140 students during the fall 2015 semester. Though it is not random, it is likely to be representative of all math 140 students in all semester.

Individuals Independent? No. The individual students came from the same classes.

Expected Counts at least 5? Yes. All of the expected counts were greater than 5.

$$\text{Degrees of freedom} = (r - 1)(c - 1) = (5 - 1)(2 - 1) = (4)(1) = 4$$

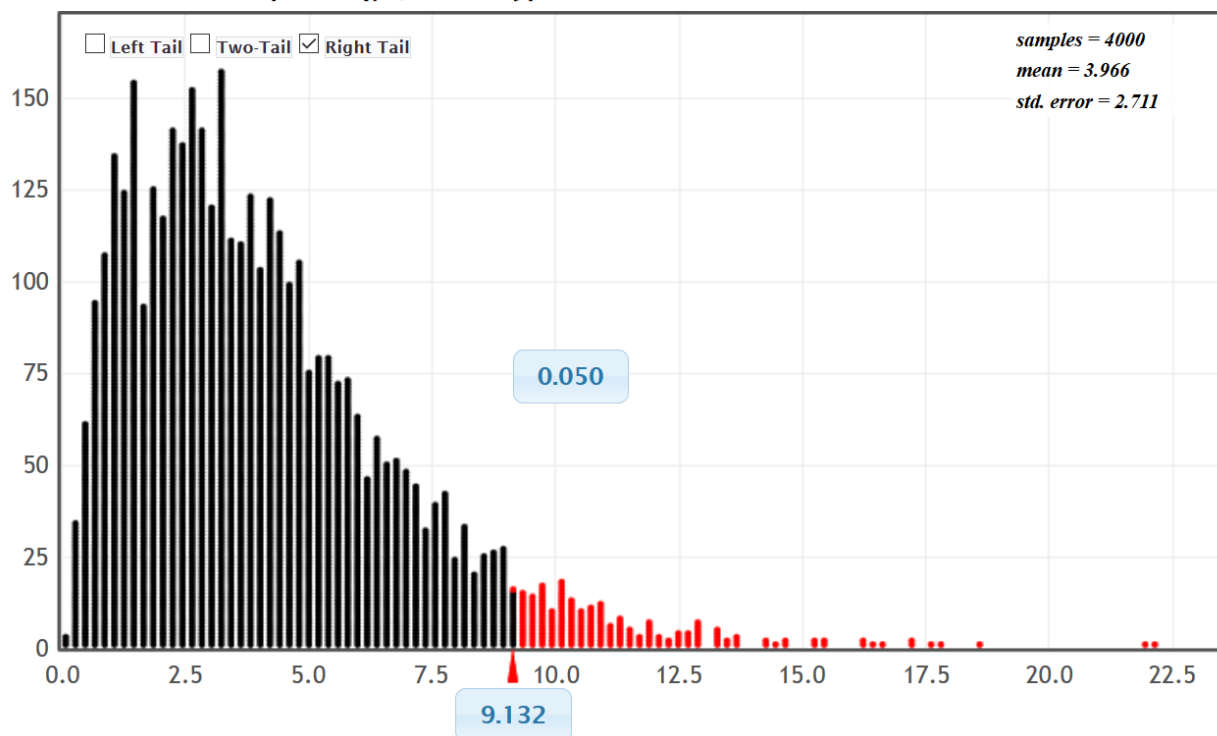
$$\chi^2 \text{ Test Statistic} = 7.531$$

The sum of the averages of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis is 7.531.



Critical Value and tail Calculation determined by significance level (simulations will vary)

Randomization Dotplot of χ^2 , Null hypothesis: No Association



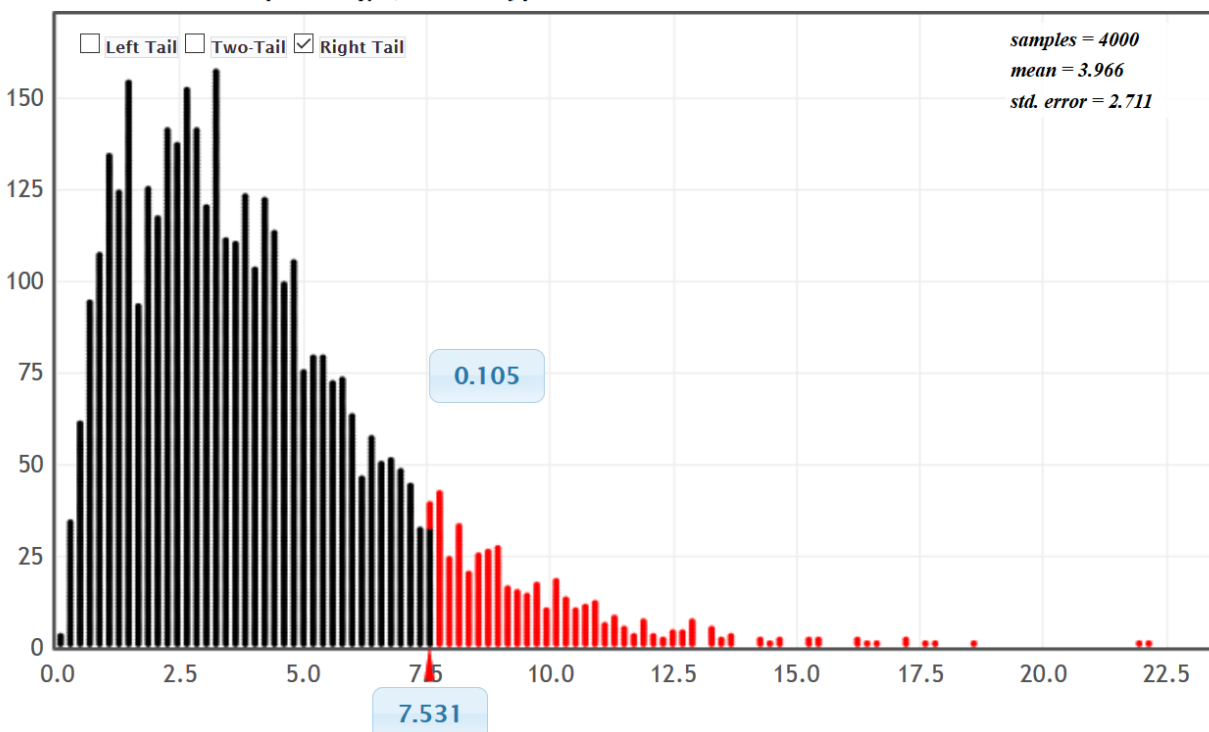
Approximate Critical value = 9.132 (answers may vary)

Notice that the Chi-square test statistic does not fall in the tail determined by the critical value. So the sample data does not significantly disagree with the null hypothesis. This also indicates that the observed sample counts were not significantly different than the expected counts.



P-value calculation with test statistic (P-values will vary)

Randomization Dotplot of χ^2 , Null hypothesis: No Association



Approximate P-value = 0.105 = 10.5% (answers will vary)

If the null hypothesis is true, there is a 10.5% probability of getting the sample data or more extreme by sampling variability.

The P-value is higher than the 5% significance level.

This sample data could have occurred by sampling variability.

Fail to reject H_0 .

Conclusion: There is NOT significant evidence to reject the claim that tattoos are not related to social media.

The P-value is high, indicating that tattoos are likely to be not related. However, the high P-value is not evidence and this data did not pass all of the assumptions for randomized simulation.

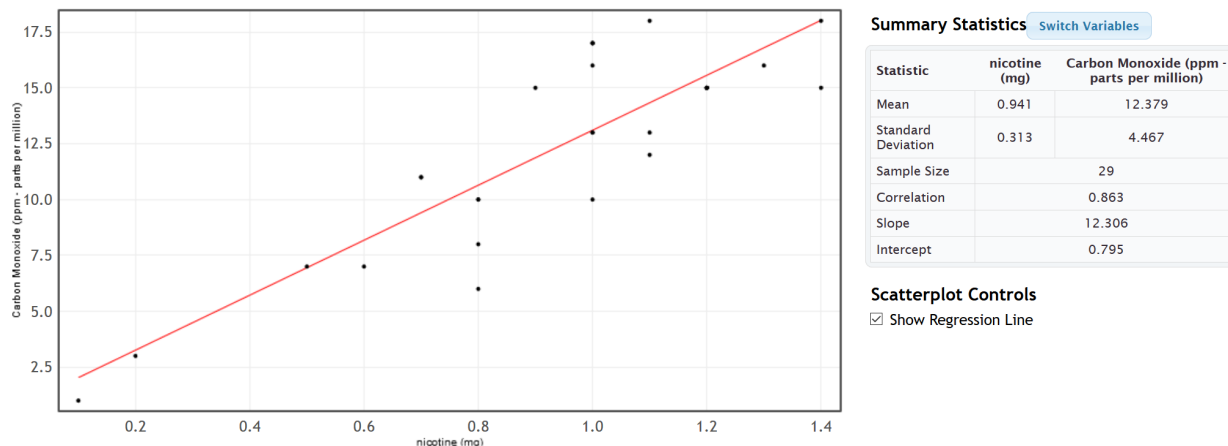
Section 4G Odd Answers

1. The response variable (Y) is the focus of the correlation study and the variable you want to make predictions about.
3. R-squared is the percentage of variability in the response variable (Y) that can be explained by the explanatory variable (X).
5. The slope of the regression line is the amount of increase or decrease in the response variable (Y) for every 1 unit increase in the explanatory variable (x).



7. Regression line formulas are only accurate in the scope of the X-values. Extrapolation is making a prediction outside the scope of the X-value. So when a person extrapolates, they plug in a number into the formula that the formula was never designed for. Extrapolation results in predictions that are not accurate and have a lot of error.

9.



a.

The scatterplot and the correlation coefficient indicate a strong positive correlation between nicotine and carbon monoxide.

b.

$$r^2 = 0.863^2 \approx 0.745 = 74.5\%$$

74.5% of the variability in carbon monoxide can be explained by the linear relationship with nicotine.

c.

$$\text{Slope} = +12.306$$

For every one mg increase in nicotine, the average carbon monoxide is increasing 12.306 ppm.

d.

$$\text{Y-intercept} = 0.795$$

If a cigarette has zero mg of nicotine, then the predicted amount of carbon monoxide would be 0.795 ppm.

Yes. The Y-intercept sentence seems to make sense. The Y-intercept is probably not very accurate since zero is not in the scope of the x values. Hence plugging in zero would be an extrapolation.

e.

The points in the scatterplot are 2.3 ppm from the regression line on average.

The average prediction error is 2.3 ppm.



f.

Regression Line: $Y = 0.795 + 12.306 X$ Prediction of Y (ppm) when $X = 1.2$ mg nicotine

$$Y = 0.795 + 12.306 X$$

$$Y = 0.795 + 12.306 (1.2)$$

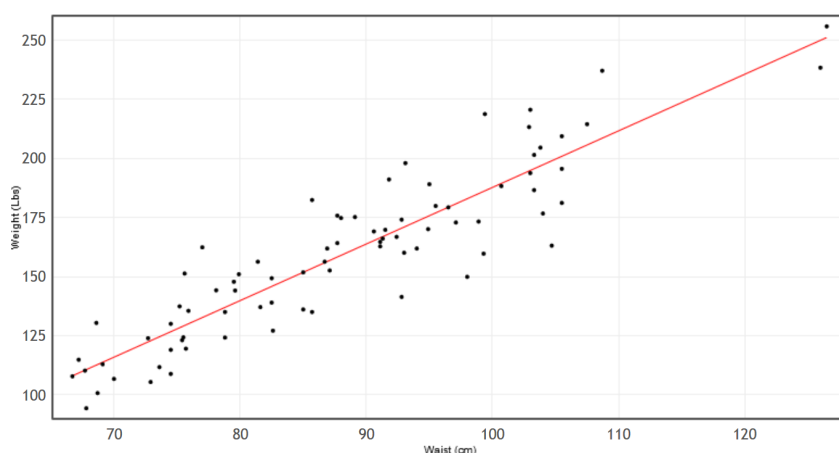
$$Y = 0.795 + 14.7672$$

$$Y = 15.5622 \text{ ppm}$$

If a cigarette has 1.2 mg of nicotine, we predict the carbon monoxide to be 15.5622 ppm.

(This prediction could be off by 2.3 ppm on average.)

11.

Summary Statistics [Switch Variables](#)

Statistic	Waist (cm)	Weight (Lbs)
Mean	88.159	159.385
Standard Deviation	13.229	34.877
Sample Size	80	
Correlation	0.908	
Slope	2.395	
Intercept	-51.728	

Scatterplot Controls

☒ Show Regression Line

a.

The scatterplot and the correlation coefficient indicate a strong positive correlation between nicotine and carbon monoxide.

b.

$$r^2 = 0.908^2 \approx 0.824 = 82.4\%$$

82.4% of the variability in weight can be explained by the linear relationship with waist size.

c.

$$\text{Slope} = +2.395$$

For every one cm increase in waist size, the average weight of the adults is increasing 2.395 pounds.

d.

$$\text{Y-intercept} = -51.728$$

If the waist size was zero cm, then the predicted weight would be -51.728 pounds.

No the Y-intercept does not make sense. It is impossible for the waist size of an adult to be zero cm. It is also impossible for the weight of an adult to be -51.728 pounds. The Y-intercept is not accurate since zero is not in the scope of the x values. Hence plugging in zero would be an extrapolation.



e.

The points in the scatterplot are 14.6809 pounds from the regression line on average.

The average prediction error is 14.6809 pounds.

f.

Regression Line: $Y = -51.728 + 2.395 X$

Prediction of Y (pounds) when X = 100 cm (waist)

$$Y = -51.728 + 2.395 X$$

$$Y = -51.728 + 2.395 (100)$$

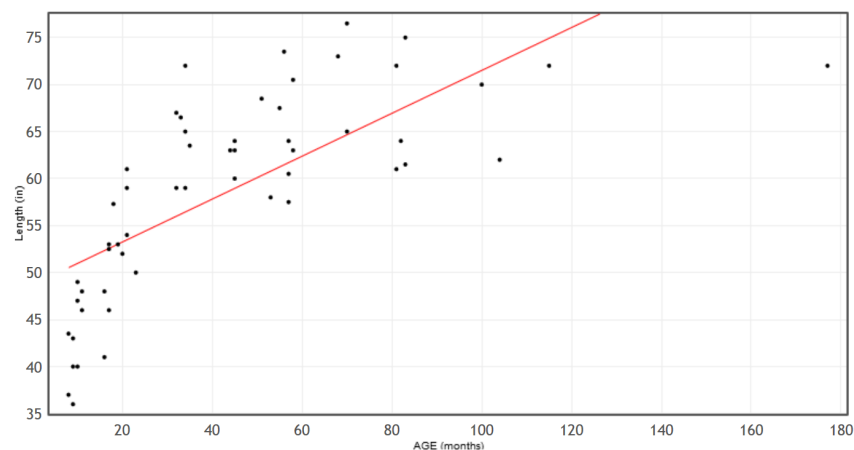
$$Y = -51.728 + 239.5$$

$$Y = 187.772 \text{ pounds}$$

If an adults waist size is 100 cm, we predict the weight to be 187.772 pounds.

(This prediction could be off by 14.608 pounds on average.)

13.



Summary Statistics [Switch Variables](#)

Statistic	AGE (months)	Length (in)
Mean	43.519	58.617
Standard Deviation	33.721	10.701
Sample Size	54	
Correlation	0.719	
Slope	0.228	
Intercept	48.69	

Scatterplot Controls

☒ Show Regression Line

a.

The scatterplot and the correlation coefficient indicate a strong positive correlation between the age and length of bears.

b.

$$r^2 = 0.719^2 \approx 0.517 = 51.7\%$$

51.7% of the variability in bear length can be explained by the linear relationship with the age of the bear.

c.

$$\text{Slope} = +0.228$$

For every one month older a bear gets, the average length of the bears is increasing 0.228 inch.



d.

Y-intercept = 48.69

If the age of the bear is zero months old, then the predicted length would be 48.69 inches.

The Y-intercept does make sense as the average length of a newborn bear. The Y-intercept is not accurate though since zero is not in the scope of the x values. Hence plugging in zero would be an extrapolation. It is no surprise that the predicted length is way off from what we expect in a newborn bear.

e.

The points in the scatterplot are 7.51 inches from the regression line on average.

The average prediction error is 7.51 inches.

f.

Regression Line: $Y = 48.69 + 0.228 X$

Prediction of Y (length in inches) when X = 24 months (age)

$$Y = 48.69 + 0.228 X$$

$$Y = 48.69 + 0.228 (24)$$

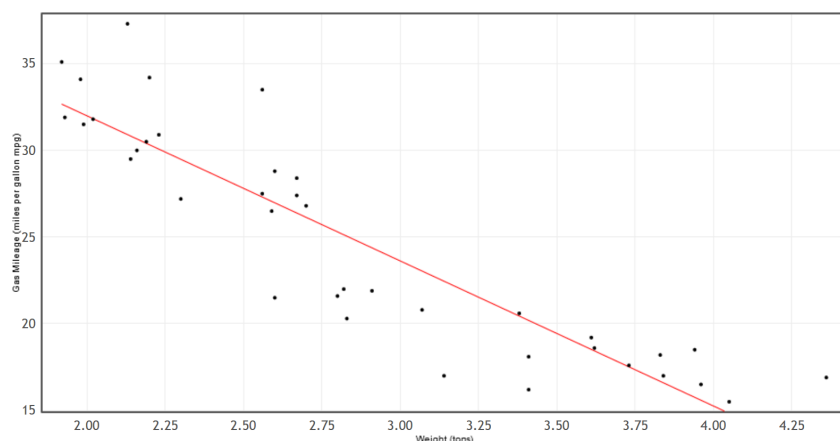
$$Y = 48.69 + 5.472$$

$$Y = 54.162 \text{ inches}$$

If a bear is 24 months old, we predict the length to be 54.162 inches.

(This prediction could be off by 7.51 inches on average.)

15.



Summary Statistics [Switch Variables](#)

Statistic	Weight (tons)	Gas Mileage (miles per gallon mpg)
Mean	2.864	24.761
Standard Deviation	0.706	6.547
Sample Size	38	
Correlation	-0.903	
Slope	-8.372	
Intercept	48.741	

Scatterplot Controls

☒ Show Regression Line

a.

The scatterplot and the correlation coefficient indicate a strong negative correlation between the weight of a car and the gas mileage.

b.

$$r^2 = (-0.903)^2 \approx 0.815 = 81.5\%$$

81.5% of the variability in gas mileage can be explained by the linear relationship with the weight of the car.



c.

Slope = -8.372

For every 1 ton increase in the weight of a car, the average gas mileage decreases 8.372 mpg.

d.

Y-intercept = 48.741

If the weight of a car is zero tons, then the predicted gas mileage would be 48.741 mpg.

The Y-intercept does not make sense. It is impossible for the weight of a car to be zero tons. The Y-intercept is not accurate though since zero is not in the scope of the x values. Hence plugging in zero would be an extrapolation.

e.

The points in the scatterplot are 2.8516 mpg from the regression line on average.

The average prediction error is 2.8516 mpg.

f.

Regression Line: $Y = 48.741 - 8.372 X$

Prediction of Y (mpg) of car when weight $X = 3$ tons

 $Y = 48.741 - 8.372 X$ $Y = 48.741 - 8.372 (3)$ $Y = 48.741 - 25.116$ $Y = 23.625$ mpg

If a car weighs 3 tons, we predict the average gas mileage to be 23.625 mpg.

(This prediction could be off by 2.8516 mpg on average.)

Section 4H Odd Answers

	T-test statistic or Correlation Coefficient (r)	Sentence to explain T-test statistic or Correlation Coefficient (r)	Critical Value (T or r)	Does the T-test statistic or r-value fall in a tail determined by a critical value? (Yes or No)	Does sample data significantly disagree with H_0 ?
1.	$T = -2.441$	The slope is 2.441 standard errors below zero.	± 1.775	Yes. In tail.	Significantly Disagrees with H_0
2.	$r = 0.183$		0.316		
3.	$T = +1.166$	The slope is 1.166 standard errors above zero.	+2.003	No. Not in tail.	Does NOT significantly Disagrees with H_0
4.	$r = -0.799$		± 0.286		
5.	$T = +3.118$	The slope is 3.118 standard errors above zero.	+2.714	Yes. In tail.	Significantly Disagrees with H_0
6.	$r = 0.921$		0.339		
7.	$T = -0.852$	The slope is 0.852 standard errors below zero.	± 2.322	No. Not in tail.	Does NOT significantly Disagrees with H_0
8.	$r = -0.026$		-0.279		



9.	$T = +1.339$	The slope is 1.339 standard errors above zero.	± 1.997	No. Not in tail.	Does NOT significantly Disagrees with H_0
10.	$r = 0.483$		+0.303		

	P-value Proportion	P-value %	Sentence to explain the P-value	Significance Level %	Significance level Proportion	If H_0 is true, could the sample data occur by random chance or is it unlikely?	Reject H_0 or Fail to reject H_0 ?
11.	0.521	52.1%	If H_0 is true, there is a 52.1% probability of getting the sample data or more extreme by sampling variability.	10%	0.10	Could be sampling variability (random chance)	Fail to reject H_0
12.	0.0426			1%			
13.	3.41×10^{-5}	0.00341%	If H_0 is true, there is a 0.00341% probability of getting the sample data or more extreme by sampling variability.	5%	0.05	Unlikely to be sampling variability (random chance)	Reject H_0
14.	0.0033			1%			
15.	0.768	76.8%	If H_0 is true, there is a 76.8% probability of getting the sample data or more extreme by sampling variability.	5%	0.05	Could be sampling variability (random chance)	Fail to reject H_0
16.	0			10%			
17.	0.0428	4.28%	If H_0 is true, there is a 4.28% probability of getting the sample data or more extreme by sampling variability.	5%	0.05	Unlikely to be sampling variability (random chance)	Reject H_0
18.	0.277			10%			
19.	6.04×10^{-6}	0.000604%	If H_0 is true, there is a 0.000604% probability of getting the sample data or more extreme by sampling variability.	1%	0.01	Unlikely to be sampling variability (random chance)	Reject H_0
20.	0.0178			5%			

21.

Correlation Test Assumptions: Quantitative ordered pair sample data collected randomly, Data values within the sample should be independent of each other, the sample size should be at least 30, the scatterplot and correlation coefficient (r) should show some linear pattern, there should be no influential outliers in the scatterplot, the histogram of the residuals should be nearly normal, the histogram of the residuals should be centered close to zero, the residual plot versus the x variables should be evenly spread out.



23.

Look for points in the scatterplot that look very far from the regression line vertically. If the correlation coefficient is close to +1 or -1, then there is strong correlation and it is unlikely that the scatterplot has influential outliers. If the correlation coefficient gets closer to zero, then there may be influential outliers.

25.

Hold your fingers horizontally and put all of the dots in the residual plots between your fingers. As you go across the plot, if your fingers remain about the same distance apart, then the graph is probably evenly spaced. If your fingers get much closer in certain parts of the graph and farther away in others, it is probably not evenly spaced.

27.

H_0 : Population Slope (β_1) = 0 (No correlation) CLAIM

H_A : Population Slope (β_1) \neq 0 (Is correlation)

Assumptions:

Quantitative ordered pair sample data collected randomly? Yes. Given

Data values within the sample should be independent of each other? Yes since it is a small random sample from a large population of all women.

The sample size should be at least 30? Yes. There are 40 women in the data.

The scatterplot and correlation coefficient (r) should show some linear pattern? Yes. The correlation coefficient and the scatterplot indicate a linear pattern. The correlation coefficient is not close to zero.

There should be no influential outliers in the scatterplot? Yes. The strong correlation coefficient and the scatterplot indicate no influential outliers.

The histogram of the residuals should be nearly normal? No. The histogram looks skewed left.

The histogram of the residuals should be centered close to zero? Maybe. The highest bar in the histogram is really close to the zero line but is a little off.

The residual plot versus the x variables should be evenly spread out? No. There is a drastic fan shape or sidewise V pattern. It is not evenly spaced.

Correlation coefficient (r) = 0.7854

There is a strong positive correlation between the systolic and diastolic blood pressure for women.

Slope = 0.5335

For every 1 mm of Hg increase in a woman's systolic blood pressure, the diastolic blood pressure increases 0.5335 mm of Hg.

T-test statistic for correlation = 7.8209

The sample slope is 7.8209 standard errors above zero.

Sample data does significantly disagree with null hypothesis since the test statistic did fall in the tail determined by the critical values.

The slope of the regression line is significantly different (higher) than zero since the test statistic did fall in the tail determined by the critical values.

P-value (answer will vary) = 0.000000019615 = 0.00000019615%

P-value is lower than 5% significance level.

Unlikely to be sampling variability.



Reject the null hypothesis.

There is significant evidence to reject the claim that there is no linear relationship (no correlation) between the systolic and diastolic blood pressures for women.

29.

H_0 : Population Slope (β_1) = 0 (No correlation) CLAIM

H_A : Population Slope (β_1) \neq 0 (Is correlation)

Assumptions:

Quantitative ordered pair sample data collected randomly? Yes. Given

Data values within the sample should be independent of each other? Maybe not. If the bears came from the same area then they would not be independent.

The sample size should be at least 30? Yes. There are 54 bears in the data.

The scatterplot and correlation coefficient (r) should show some linear pattern? Yes. The correlation coefficient and the scatterplot indicate a linear pattern. The correlation coefficient is not close to zero.

There should be no influential outliers in the scatterplot? Yes. The strong correlation coefficient and the scatterplot indicate no influential outliers.

The histogram of the residuals should be nearly normal? No. The histogram looks skewed right.

The histogram of the residuals should be centered close to zero? Maybe. The highest bar in the histogram is really close to the zero line but is a little off.

The residual plot verses the x variables should be evenly spread out? No. There is a drastic fan shape or sidewise V pattern. It is not evenly spaced.

Correlation coefficient (r) = 0.9341

There is a strong positive correlation between the neck circumference and weight of the bears in the sample.

Slope = 20.1694

For every 1 inch increase in the bears neck circumference, the weight increases 20.1694 pounds.

T-test statistic for correlation = 18.8612

The sample slope is 18.8612 standard errors above zero.

Sample data does significantly disagree with null hypothesis since the test statistic did fall in the tail determined by the critical values.

The slope of the regression line is significantly different (higher) than zero since the test statistic did fall in the tail determined by the critical values.

P-value (answer will vary) = 0 = 0%

P-value is lower than 5% significance level.

Unlikely to be sampling variability.

Reject the null hypothesis.

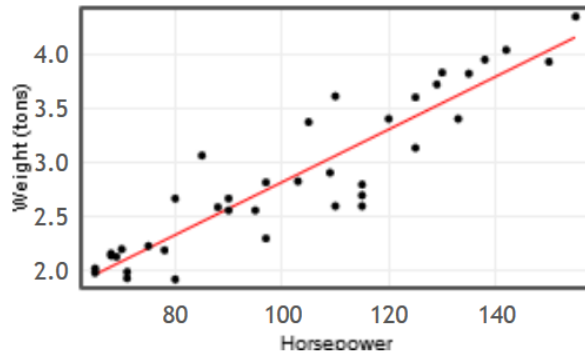
There is significant evidence to reject the claim that there is no linear relationship (no correlation) between the neck circumference and weight of bears.



31.

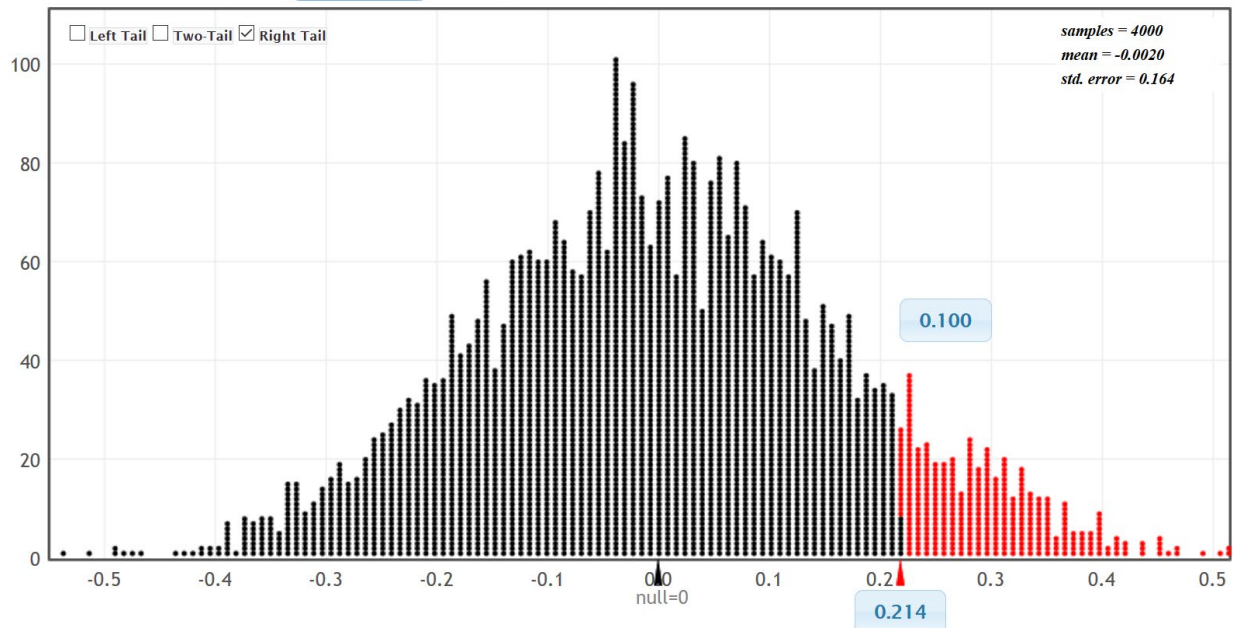
Original Sample

$n = 38$, $r = 0.917$, $\text{slope} = +0.024$, $\text{intercept} = +0.372$



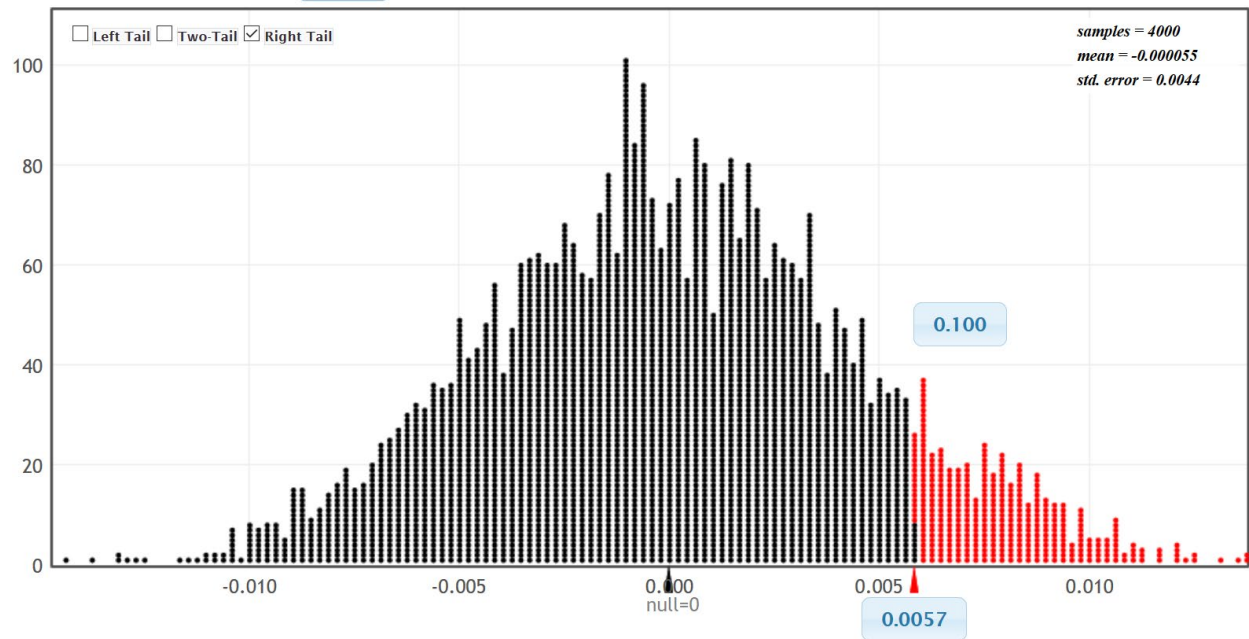
Tail for Correlation Coefficient (right tail and 10% sig level) (simulations will vary)

Randomization Dotplot of Correlation Null hypothesis: $\rho = 0$



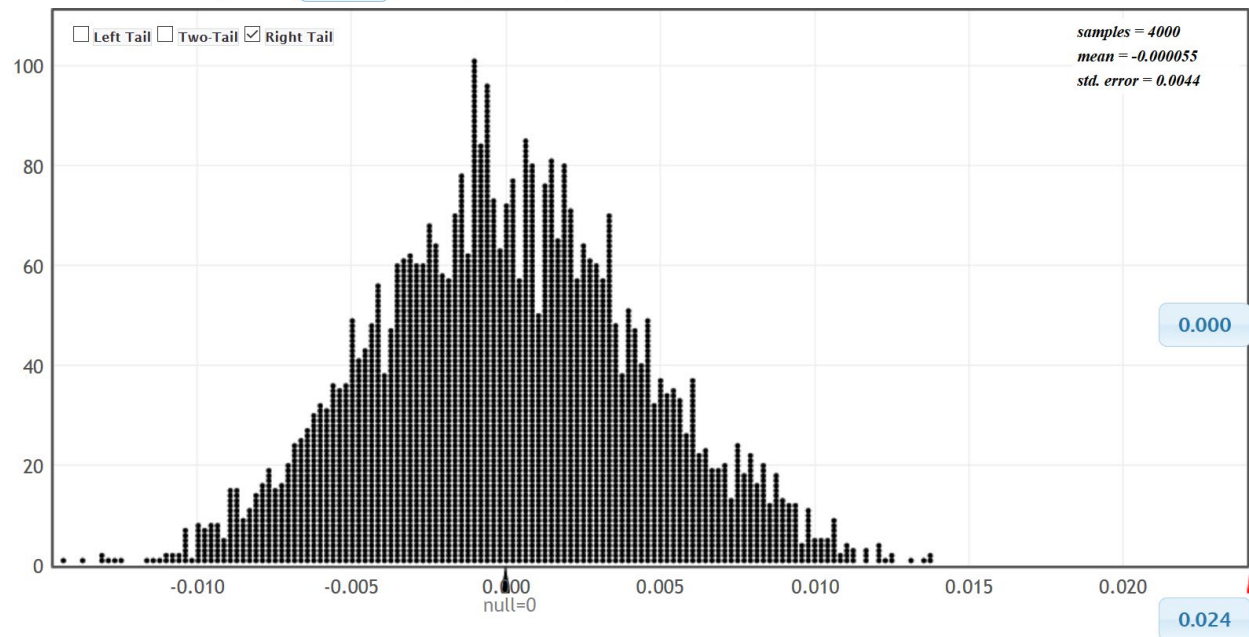
Tail for Slope (right tail and 10% sig level) (simulations will vary)

Randomization Dotplot of **Slope** Null hypothesis: $\beta_1 = 0$



P-value calculation from slope simulation (answers will vary)

Randomization Dotplot of **Slope** Null hypothesis: $\beta_1 = 0$



H_0 : Population Slope (β_1) = 0 (No correlation)

H_A : Population Slope (β_1) > 0 (Is positive correlation) CLAIM

The correlation coefficient (r) = 0.917

There is a strong positive correlation between the horsepower and weights of cars.



The correlation coefficient does fall in the tail determined by the correlation simulation and the 10% sig level.

The slope = 0.024

For every 1 horsepower increase in a car, the weights are increasing 0.024 tons on average.

The slope of 0.024 does fall in the tail determined by the slope simulation and the 10% sig level.

Sample data does significantly disagree with null hypothesis since the slope and correlation coefficient did fall in the tail determined by the simulations.

The slope of the regression line is significantly different (higher) than zero since the slope did fall in the tail determined by the simulation.

P-value (answer will vary) = 0 = 0%

P-value is lower than 10% significance level.

Unlikely to be sampling variability.

Reject the null hypothesis.

There is significant evidence to support the claim that there is a positive linear relationship (positive correlation) between the horsepower and weight of cars.

T-test statistic calculation (answers will vary) = $(0.024 - 0) \div 0.0044 = 5.4545$

The slope is 5.4545 standard errors above zero.

Chapter 4 Review All Answers

1.

Correlation: Statistical analysis that determines if there is a relationship between two different quantitative variables.

Regression: Statistical analysis that involves finding the line or model that best fits a quantitative relationship, using the model to make predictions, and analyzing error in those predictions.

Explanatory Variable (x): Another name for the x-variable or independent variable in a correlation study.

Response Variable (y): Another name for the y-variable or dependent variable in a correlation study.

Correlation Coefficient (r): A statistic between -1 and $+1$ that measures the strength and direction of linear relationships between two quantitative variables.

R-squared (r^2): Also called the coefficient of determination. This statistic measures the percent of variability in the y-variable that can be explained by the linear relationship with the x-variable.

Residual ($y - \hat{y}$): The vertical distance between the regression line and a point in the scatterplot.

Standard Deviation of the Residual Errors (s_e): A statistic that measures how far points in a scatterplot are from the regression line on average and measures the average amount of prediction error.

Slope (b_1): The amount of increase or decrease in the y-variable for every one-unit increase in the x-variable.

Y-Intercept (b_0): The predicted y-value when the x-value is zero.

Regression Line ($\hat{y} = b_0 + b_1x$): Also called the line of best fit or the line of least squares. This line minimizes the vertical distances between it and all the points in the scatterplot.



Scatterplot: A graph for visualizing the relationship between two quantitative ordered pair variables. The ordered pairs (x, y) are plotted on the rectangular coordinate system.

Residual Plot: A graph that pairs the residuals with the x values. This graph should be evenly spread out and not fan shaped.

Histogram of the Residuals: A graph showing the shape of the residuals. This graph should be nearly normal and centered close to zero.

Hypothesis Test: A procedure for testing a claim about a population.

Random Chance: Another word for sampling variability. The principle that random samples from the same population will usually be different and give very different statistics.

Critical Value: If the test statistic is higher than this number, then the sample data significantly disagrees with the null hypothesis. The z or t score critical values are also used to calculate margin of error in confidence intervals.

P-value: The probability of getting the sample data or more extreme by random chance if the null hypothesis is true.

Significance Level (α): Also called the Alpha Level. If the P-value is lower than this number, then the sample data significantly disagrees with the null hypothesis and is unlikely to have happened by random chance. This is also the probability of making a type 1 error.

Randomized Simulation: A technique for visualizing sampling variability in a hypothesis test. The computer assumes the null hypothesis is true, and then generates random samples. If the sample data or test statistic falls in the tail, then the sample data significantly disagrees with the null hypothesis. This technique can also calculate the P-value without a formula.

Chi-square test statistic (χ^2): The sum of the average of the squares of the differences between the observed sample counts and the expected counts from the null hypothesis.

F-test statistic: The ratio of the variance between the groups to the variance within the groups.

T-test statistic for correlation: The number of standard errors that the slope is above or below zero.

Z-test statistic for two-population proportion: The number of standard errors that the sample proportion from group 1 is above or below the sample proportion from group 2.

T-test statistic for two-population mean: The number of standard errors that the sample mean from group 1 is above or below the sample mean from group 2.

#2-4.

Two-population proportion test

$$H_0: \pi_1 = \pi_2$$

$$H_A: \pi_1 > \pi_2$$

OR

$$H_0: \pi_1 - \pi_2 = 0$$

$$H_A: \pi_1 - \pi_2 > 0$$

Z-test statistic

Assumptions: Random Samples, Individuals within and between the samples are independent, both samples have at least 10 success and at least 10 failures.



Goodness of Fit proportion test

$$H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4$$

$$H_A: \text{at least one } \neq$$

OR

$$H_0: \pi_1 = 0.5, \pi_2 = 0.25, \pi_3 = 0.15, \pi_4 = 0.1$$

$$H_A: \text{at least one } \neq$$

Chi-square test statistic (χ^2)

Assumptions: Random Samples, Individuals within and between the samples are independent, all expected counts are at least 5.

Categorical Association Test

$$H_0: \text{Categorical variables are not related.}$$

$$H_A: \text{Categorical variables are related.}$$

Chi-square test statistic (χ^2)

Assumptions: Random Sample or Random Samples, Individuals within (and between) the samples are independent, all expected counts are at least 5.

Two-population mean test

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 < \mu_2$$

OR

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 < 0$$

OR

$$H_0: \mu_d = 0$$

$$H_A: \mu_d < 0$$

T-test statistic for two-population mean.

Assumptions: Random Samples, Individuals within (and between) the samples are independent, both samples have a sample size of at least 30 or nearly normal.

ANOVA test

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A: \text{at least one } \neq$$

F-test statistic

Assumptions: Random Samples, Individuals within (and between) the samples are independent, all samples have a sample size of at least 30 or nearly normal, sample standard deviations are close.

Correlation Test

$$H_0: \text{Population Slope } (\beta_1) = 0$$



H_A : Population Slope (β_1) $\neq 0$

OR

H_0 : Population Correlation Coefficient (ρ) = 0

H_A : Population Correlation Coefficient (ρ) $\neq 0$

T-test statistic for correlation.

Assumptions: Quantitative ordered pair sample data collected randomly, Data values within the sample should be independent of each other, the sample size should be at least 30, the scatterplot and correlation coefficient (r) should show some linear pattern, there should be no influential outliers in the scatterplot, the histogram of the residuals should be nearly normal, the histogram of the residuals should be centered close to zero, the residual plot verses the x variables should be evenly spread out.

5.

Test Statistic	Critical Value	Does the sample data significantly disagree with H_0 ?	Explain why.
F = 2.174	3.823	Not significantly disagree.	Test stat not in tail.
T = -2.556	± 1.96	Significantly disagree.	Test stat in tail.
$\chi^2 = 16.87$	9.977	Significantly disagree.	Test stat in tail.
F = 5.339	2.742	Significantly disagree.	Test stat in tail.
T = 1.349	± 2.576	Not significantly disagree.	Test stat not in tail.
$\chi^2 = 1.883$	7.187	Not significantly disagree.	Test stat not in tail.

6.

P-value	P-value %	Significance Level	Does the sample data significantly disagree with H_0 ?	Could be random chance or Unlikely?	Reject H_0 or fail to reject?
0.238	23.8%	5%	Not Significantly disagree.	Could be.	Fail to reject H_0
0.0003	0.03%	1%	Significantly disagree.	Unlikely	Reject H_0
5.7×10^{-6}	0.00057%	10%	Significantly disagree.	Unlikely	Reject H_0
0.441	44.1%	5%	Not Significantly disagree.	Could be.	Fail to reject H_0
0.138	13.8%	1%	Not Significantly disagree.	Could be.	Fail to reject H_0
0	0%	10%	Significantly disagree.	Unlikely	Reject H_0

7.

P-value	Sig Level	Claim	Conclusion
0.238	5%	H_0	There is not significant evidence to reject the claim.
0.0003	1%	H_A	There is significant evidence to support the claim.
5.7×10^{-6}	10%	H_0	There is significant evidence to reject the claim.
0.441	5%	H_A	There is not significant evidence to support the claim.
0.138	1%	H_0	There is not significant evidence to reject the claim.
0	10%	H_A	There is significant evidence to support the claim.



8.

Correlation Test H_0 : Population Slope (β_1) = 0 H_A : Population Slope (β_1) \neq 0

OR

 H_0 : Population Correlation Coefficient (ρ) = 0 H_A : Population Correlation Coefficient (ρ) \neq 0

T-test statistic for correlation.

Assumptions: Quantitative ordered pair sample data collected randomly, Data values within the sample should be independent of each other, the sample size should be at least 30, the scatterplot and correlation coefficient (r) should show some linear pattern, there should be no influential outliers in the scatterplot, the histogram of the residuals should be nearly normal, the histogram of the residuals should be centered close to zero, the residual plot verses the x variables should be evenly spread out.

9.

Categorical Association Test H_0 : Categorical variables are not related. H_A : Categorical variables are related.Chi-square test statistic (χ^2)

Assumptions: Random Sample or Random Samples, Individuals within (and between) the samples are independent, all expected counts are at least 5.

10.

ANOVA test H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$ H_A : at least one \neq

F-test statistic

Assumptions: Random Samples, Individuals within (and between) the samples are independent, all samples have a sample size of at least 30 or nearly normal, sample standard deviations are close.

11.

Goodness of Fit proportion test H_0 : $\pi_1 = \pi_2 = \pi_3 = \pi_4$ H_A : at least one \neq

OR

 H_0 : $\pi_1 = 0.5, \pi_2 = 0.25, \pi_3 = 0.15, \pi_4 = 0.1$ H_A : at least one \neq Chi-square test statistic (χ^2)

Assumptions: Random Samples, Individuals within and between the samples are independent, all expected counts are at least 5.



12.

ANOVA since the amount of money is quantitative and the city is categorical.

13.

Goodness of Fit since we are checking a specific percentage in multiple groups.

14.

Correlation test since the amount of rainfall and number of fires are both quantitative variables.

15.

Categorical Association Test since the type of health insurance and the education level are both categorical variables with multiple responses.

16.

% of knee injuries from soccer = $31/100 = 31\%$

% of knee injuries from tennis = $5/100 = 5\%$

These percentages are significantly different indicating the sport may be related to having a knee injury.

Goodness of Fit test.

$H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6$ (Claim)

H_A : at least one \neq

Assumptions:

Random? Yes. Given

Individuals independent? Yes since it is a small random sample from a large population.

Expected counts at least 5? Yes. All expected counts are 16.7.

Chi-square test statistic = 28.16

Sample data significantly disagrees with null hypothesis since the test statistic falls in the tail determined by the critical value.

Observed sample counts significantly disagrees with expected counts in the null hypothesis since the test statistic falls in the tail determined by the critical value.

P-value = 0.0033869%

P-value is less than 1% significance level.

Unlikely to be sampling variability.

Reject the null hypothesis.

There is significant evidence to reject the claim that the percentage of knee injuries are the same in the various sports.

The data indicates that having a knee injury is related to the sport.



17.

% of raccoonss have rabies = $7/27 = 25.9\%$

% of squirrels have rabies = $17/38 = 44.7\%$

% of chipmunks have rabies = $8/30 = 26.7\%$

If these percentages are close, it would indicate that the type of animal is not related to having rabies.

If these percentages are significantly different, it would indicate that the type of animal is related to having rabies.

Categorical Association Test. (Since this data was collected from 1 random sample, it is sometimes referred to as "independence".)

H_0 : Rabies status is not related to the type of animal.

H_A : Rabies status is related to the type of animal. (Claim)

Assumptions:

Random? Yes. Given

Individuals independent? It might not be independent if all the animals came from the same region.

Expected counts at least 5? Yes. The expected counts were 12.8, 10.11, 9.09, 25.2, 19.89, and 17.91.

Chi-square test statistic = 3.467

Sample data does NOT significantly disagrees with null hypothesis since the test statistic did NOT fall in the tail determined by the critical value.

Observed sample counts do NOT significantly disagree with expected counts in the null hypothesis since the test statistic did NOT fall in the tail determined by the critical value.

P-value = 17.67%

P-value is higher than 5% significance level.

Could be sampling variability.

Fail to reject the null hypothesis.

There is not significant evidence to support the claim that the type of animal is related to rabies status.

The data indicates that the type of animal may not be related to having rabies.



18.

Original Sample[Show Details](#) $n = 2625$, $\chi^2 = 7.989$

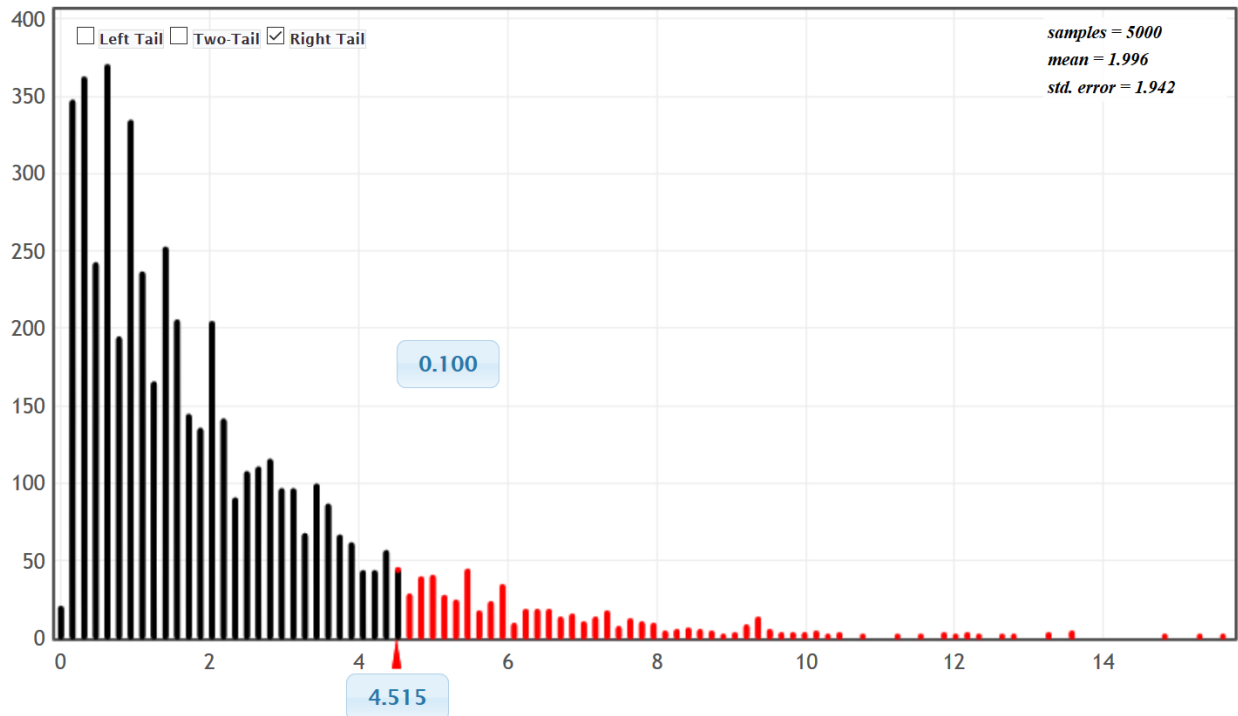
	Male	Female	Total
Agree	372	363	735
Disagree	807	1005	1812
Don't Know	34	44	78
Total	1213	1412	2625

Detailed Sample Table[Close](#)[Help](#)

	Male	Female	Total
Agree	372 339.6 3.083	363 395.4 2.649	735
Disagree	807 837.3 1.098	1005 974.7 0.943	1812
Don't Know	34 36 0.116	44 42 0.1	78
Total	1213	1412	2625

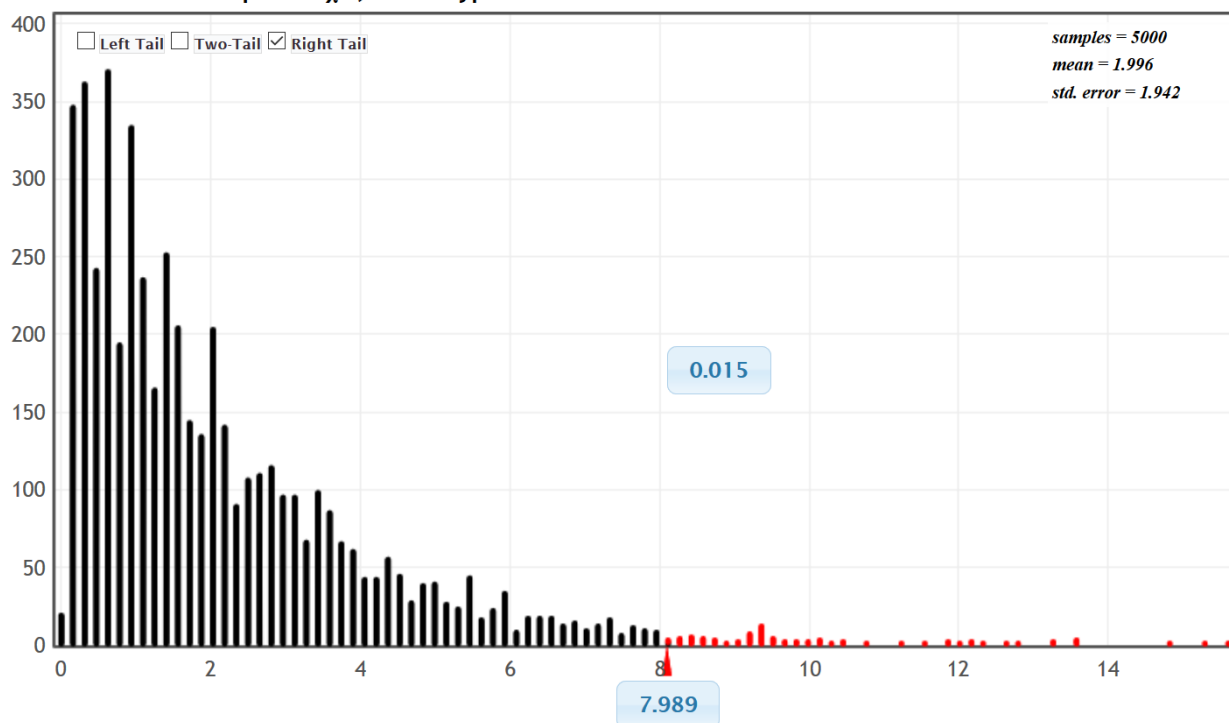
Observed, Expected, Contribution to χ^2

Critical Value Calculations (Answers will vary.)

Randomization Dotplot of χ^2 , Null hypothesis: No Association

P-value Calculation (Answers will vary.)

Randomization Dotplot of χ^2 , Null hypothesis: No Association



% of males that believe in one true love = $372/1213 = 30.7\%$

% of females that believe in one true love = $363/1412 = 25.7\%$

If these percentages are close, it would indicate that gender is not related to believing in one true love.

If these percentages are significantly different, it would indicate that gender is related to believing in one true love.

Categorical Association Test.

H_0 : Belief in one true love is not related to gender. (Claim)

H_A : Belief in one true love is related to gender.

Assumptions:

Random? Unknown.

Individuals independent? If it is a random sample, then it probably is independent.

Expected counts at least 5? Yes. The expected counts were 339.6, 395.4, 837.3, 974.7, 36, and 42.

Chi-square test statistic = 7.989

Critical value (answer will vary) = 4.515

Sample data does significantly disagree with null hypothesis since the test statistic did fall in the tail determined by the critical value.

Observed sample counts do significantly disagree with expected counts in the null hypothesis since the test statistic did fall in the tail determined by the critical value.

P-value (answer will vary) = 0.015 = 1.5%



P-value is lower than 10% significance level.

Unlikely to be sampling variability.

Reject the null hypothesis.

There is significant evidence to reject the claim that gender is not related to believing in one true love.

The data indicates that gender is related to believing in one true love.

19.

Original Sample

ANOVA Table

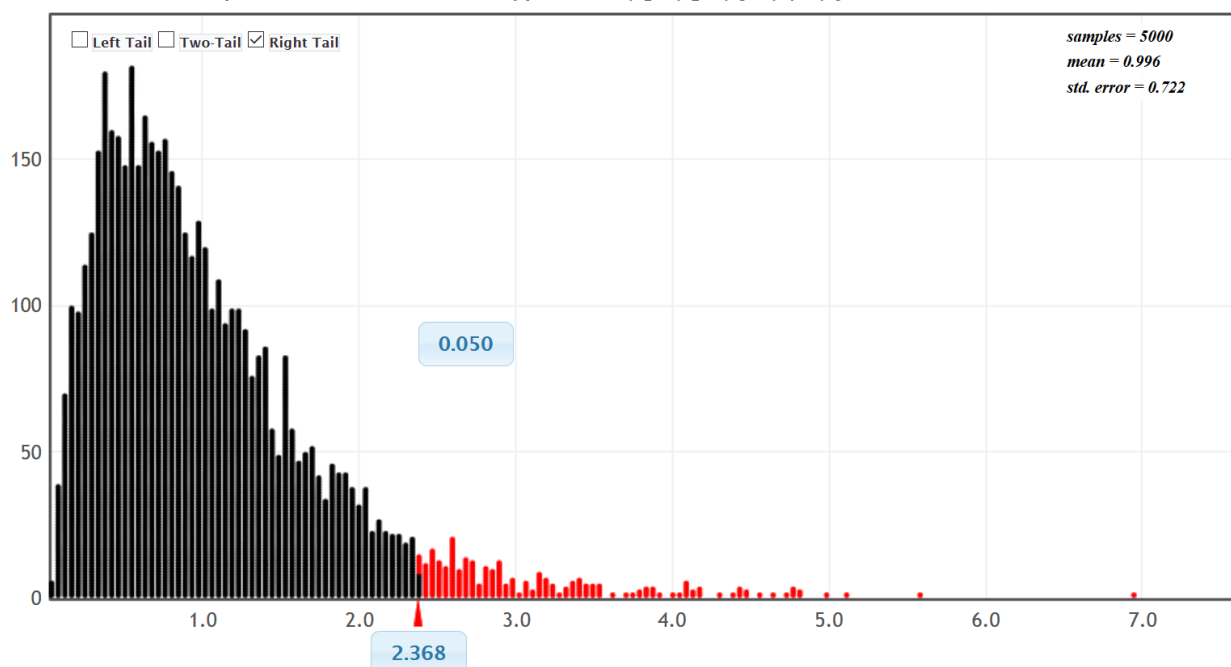
$n = 323$, $F = 0.437$

Statistics	Snapchat	Other	Facebook	Instagram	Twitter	Overall
Sample Size	71	27	72	124	29	323
Mean	11.5	10.9	12.5	11.9	11.8	11.9
Standard Deviation	6.4	5.3	8.0	5.0	3.7	6.0

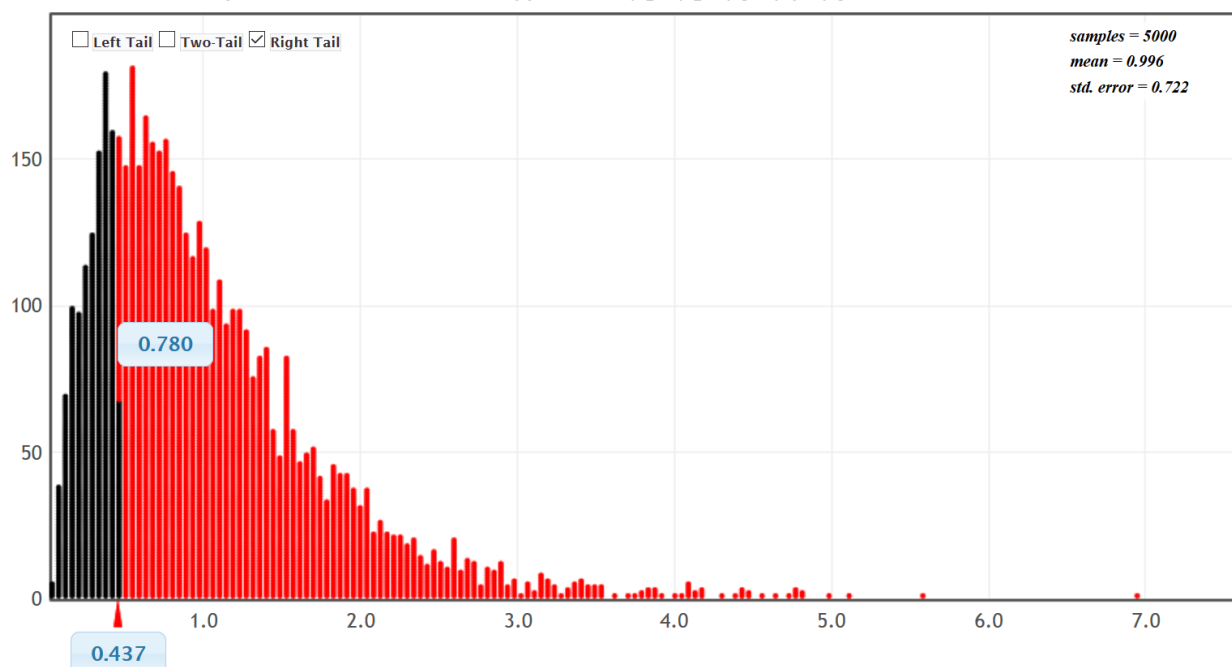
ANOVA Table

	df	SS	MS	F
Groups	4	64.48	16.12	0.437
Error	318	11720.01	36.85	
Total	322	11784.49		

Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$



Randomization Dotplot of F-statistic , Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$



H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (not related) CLAIM

H_A : at least one \neq (related)

Assumptions:

Random Samples? No. This data was a census. It may be representative though.

Individuals within (and between) the samples are independent? No. These students came from the same math 140 classes.

All samples have a sample size of at least 30 or nearly normal? Maybe. Shapes are unknown. All the sample sizes are above 30 except for twitter which was 29.

Sample standard deviations are close? No. One of the standard deviations (8) was more than twice as large as another (3.9).

F-test statistic = 0.437

Critical Value (answer will vary) = 2.368

Sample data does NOT significantly disagree with null hypothesis since the test statistic did NOT fall in the tail determined by the critical value.

The variance between the groups is NOT significantly higher than the variance within since the test statistic did NOT fall in the tail determined by the critical value.

P-value (answer will vary) = 0.780 = 78.0%

P-value is higher than 5% significance level.

Could be sampling variability.

Fail to reject the null hypothesis.



There is not significant evidence to reject the claim that the a persons favorite social media is not related to the money spent on meals.

The data indicates that social media might be NOT related to the money spent on meals. We do not have evidence and the data did not pass the assumptions.

20.

$$r = 0.7997$$

There is a strong positive correlation between the weights and BMI for the men in the sample.

$$r^2 = 0.6395 = 63.95\%$$

63.95% of the variability in body mass index can be explained by the relationship with weight.

$$\text{Slope} = 0.1042$$

For every 1 pound increase in weight, the men's BMI is increasing 0.1042 kg/m².

Y-intercept = 8.0169 (*Does not make sense since it is impossible for a mans weight to be zero.*)

If a man's weight is zero pounds, the predicted BMI would be 8.0169 kg/m².

Standard Deviation of Residual Errors = 2.0869 kg/m².

The average prediction error is 2.0869 kg/m².

The average vertical distance from the regression line is 2.0869 kg/m².

Prediction for man weighing 185 pounds.

$$\hat{y} = 8.0169 + 0.1042X$$

$$\hat{y} = 8.0169 + 0.1042(185)$$

$$\hat{y} = 8.0169 + 19.277$$

$$\hat{y} = 27.2939 \text{ kg/m}^2$$

(This prediction could have an average error of 2.0869 kg/m².)

Correlation Test

H_0 : Population Slope (β_1) = 0 (No correlation)

H_A : Population Slope (β_1) \neq 0 (Is correlation) CLAIM

OR

H_0 : Population Correlation Coefficient (ρ) = 0 (No correlation)

H_A : Population Correlation Coefficient (ρ) \neq 0 (Is correlation) CLAIM

Assumptions:

Quantitative ordered pair sample data collected randomly? Yes. Given

Data values within the sample should be independent of each other? Yes since it is a small random sample from a large population of all men.

The sample size should be at least 30? Yes. There are 40 men in the data.

The scatterplot and correlation coefficient (r) should show some linear pattern? Yes. The correlation coefficient and the scatterplot indicate a linear pattern. The correlation coefficient is not close to zero.



There should be no influential outliers in the scatterplot? Yes. The strong correlation coefficient and the scatterplot indicate no influential outliers.

The histogram of the residuals should be nearly normal? Yes. The histogram looks slightly skewed right, but can probably be thought of as nearly normal.

The histogram of the residuals should be centered close to zero? Yes. The highest bar in the histogram is touching the zero line.

The residual plot verses the x variables should be evenly spread out? Yes. There does not appear to be a drastic fan shape or sidewise V pattern. It is not perfectly even, but is close.

T-test statistic for correlation = 8.2095

Sample data does significantly disagree with null hypothesis since the test statistic did fall in the tail determined by the critical values.

The slope of the regression line is significantly different (higher) than zero since the test statistic did fall in the tail determined by the critical values.

P-value (answer will vary) = 0.00000000060619 = 0.000000060619%

P-value is lower than 1% significance level.

Unlikely to be sampling variability.

Reject the null hypothesis.

There is significant evidence to support the claim that there is a linear relationship (correlation) between the weight and BMI for men.

